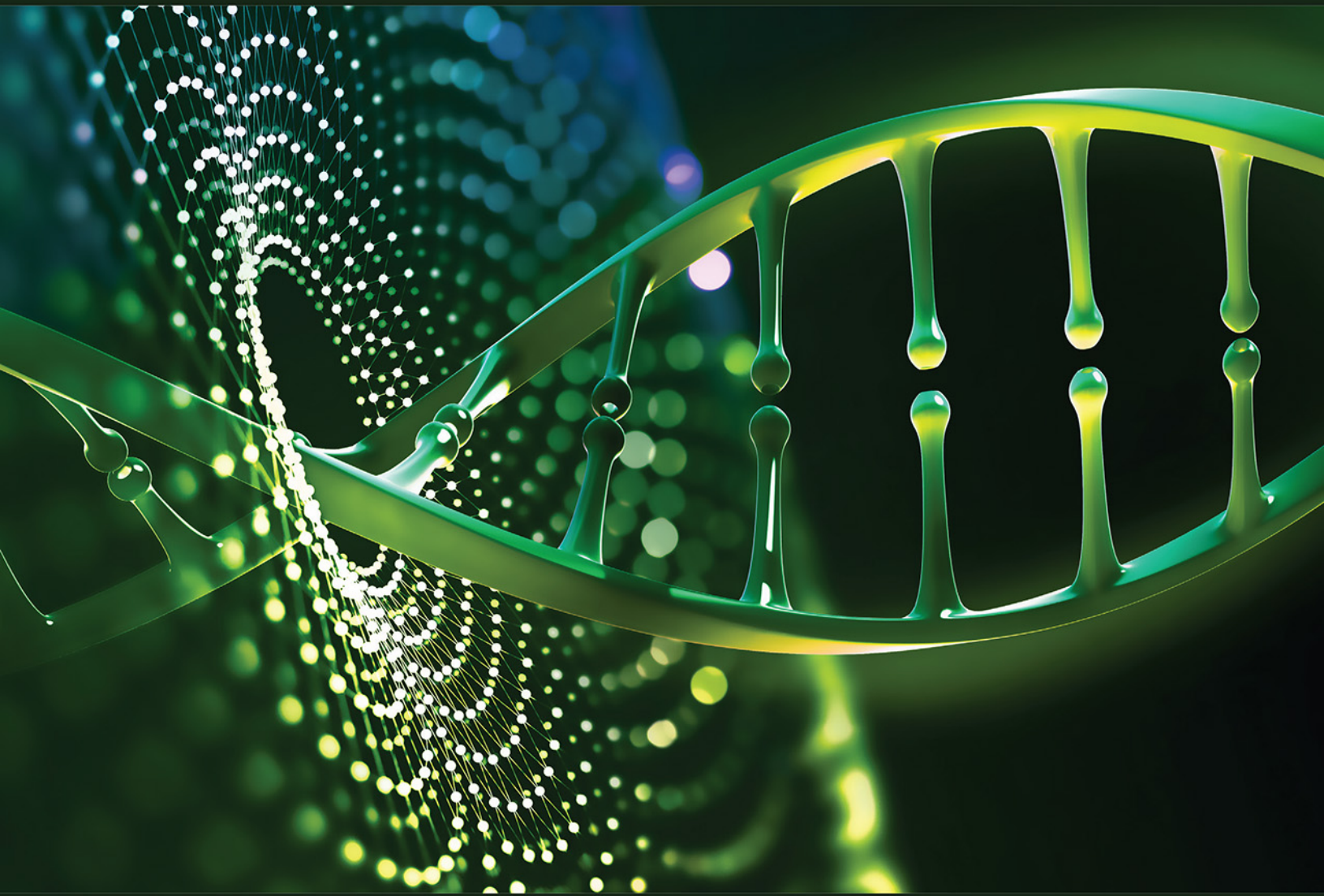


Bioinformatics in Agriculture

Next-Generation Sequencing Era



Edited by
**Pradeep Sharma, Dinesh Yadav
and Rajarshi Kumar Gaur**



Bioinformatics in Agriculture

Next-Generation Sequencing Era

This page intentionally left blank

Bioinformatics in Agriculture

Next-Generation Sequencing Era

Edited by

Pradeep Sharma

ICAR-Indian Institute of Wheat and Barley Research, Karnal, Haryana, India

Dinesh Yadav

Department of Biotechnology, Deen Dayal Upadhyaya Gorakhpur University,
Gorakhpur, Uttar Pradesh, India

Rajarshi Kumar Gaur

Department of Biotechnology, Deen Dayal Upadhyaya Gorakhpur University,
Gorakhpur, Uttar Pradesh, India



ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2022 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-323-89778-5

For Information on all Academic Press publications
visit our website at <https://www.elsevier.com/books-and-journals>

Publisher: Charlotte Cockle
Acquisitions Editor: Nancy J. Maragioglio
Editorial Project Manager: Emerald Li
Production Project Manager: R. Vijay Bharath
Cover Designer: Greg Harris

Typeset by MPS Limited, Chennai, India



Contents

List of contributors	xvii
About the editors	xxiii
Foreword	xxv
Preface	xxvii

Section I Bioinformatics and next generation sequencing technologies

1. Advances in agricultural bioinformatics: an outlook of multi “omics” approaches	3
<i>Nisha Singh, Megha Ujinwal and Anuradha Singh</i>	
1.1 Introduction	3
1.2 Different types of “omics” approaches	3
1.2.1 Phenomics	3
1.2.2 Genomics	6
1.2.3 Transcriptomics	8
1.2.4 Proteomics	10
1.2.5 Metabolomics	12
1.2.6 Ionomics	14
1.2.7 Computomics	15
1.3 Conclusions and future prospective	16
References	16
2. Promises and benefits of omics approaches to data-driven science industries	23
<i>Niranjani Iyer</i>	
2.1 Sequencing technologies	23
2.2 Advances in genome assembly technology	24
2.2.1 Algorithms in reference-based and de novo assembly	24
2.2.2 Postassembly algorithms for encoding the biology	25
2.2.3 Genome-wide association, a valuable tool mapping associations with a phenotype	26

2.3 Transcriptomics—where genome connects to gene function	26
2.3.1 Methodologies and algorithms	26
2.3.2 Noncoding RNA	28
2.3.3 Epigenomics	29
2.4 Beyond genomics and transcriptomics toward proteomics and metabolomics	29
2.4.1 Proteomics	29
2.4.2 Metabolomics	30
2.5 Integrating omics datasets	30
2.6 Challenges	31
2.7 Machine learning in omics	31
2.7.1 Machine learning for genomic studies	32
2.8 Big data storage and management	32
2.9 Future directions	33
References	33
3. Bioinformatics intervention in functional genomics: current status and future perspective—an overview	37
<i>Swati Sharma, Ashwani Kumar, Dinesh Yadav and Manoj Kumar Yadav</i>	
3.1 Introduction	37
3.2 Functional genomic approaches	37
3.3 Serial analysis of gene expression	38
3.3.1 Advantages of serial analysis of gene expression	39
3.3.2 Drawbacks of serial analysis of gene expression technique	39
3.4 DNA microarray	40
3.4.1 Applications of microarray	40
3.4.2 Drawbacks of microarray	40
3.4.3 Bioinformatics tools for microarray data analysis	41
3.5 Next-generation sequencing technologies	41
3.5.1 Illumina sequencing	41
3.5.2 Applications of next-generation sequencing	42
3.5.3 Bioinformatics tools for next-generation sequencing	43

3.6 Databases and genome annotation	43	5.3.4 Gene trait association analysis using natural diverse populations	68
3.6.1 Biological databases	43	5.3.5 Genetic transformations	68
3.6.2 Functional genomic databases	44	5.4 Genome analysis	68
3.7 Conclusion	45	5.4.1 Sequencing	69
References	46	5.4.2 Assembly	70
4. Genome informatics: present status and future prospects in agriculture	47	5.4.3 Annotation	71
<i>Pramod Kumar Yadav, Rahul Singh Jasrotia and Akanksha Jaiswar</i>		5.5 Applications of genomics	71
4.1 Introduction	47	5.5.1 Genomics in medicine	71
4.2 The evolution of DNA-seq	48	5.5.2 Genomics in synthetic biology and bioengineering	71
4.2.1 The first generation of sequencing technologies	48	5.5.3 Conservation genomics	71
4.2.2 The second generation of sequencing technologies	48	5.6 Next-generation genomics for crop improvement	71
4.2.3 The third generation of sequencing technologies	49	5.7 Genomic features for future breeding	75
4.3 Genomics in agriculture	49	References	75
4.3.1 Genome assembly	49	6. Genome-wide predictions, structural and functional annotations of plant transcription factor gene families: a bioinformatics approach	79
4.3.2 RNA-seq in agriculture	52	<i>Sudhanshu Srivastava, Kapil Gupta, Kanchan Yadav, Manoj Kumar Yadav and Dinesh Yadav</i>	
4.3.3 Databases and prediction servers	54	6.1 Transcription factor: an introduction	79
4.4 Conclusion, applications, and future prospects of next-generation sequencing in agriculture	55	6.2 Plant transcription factors and its multifarious applications	79
References	56	6.2.1 AP2/ERF family	80
5. Genomics and its role in crop improvement	61	6.2.2 bHLH family	81
<i>Ujjawal Kumar Singh Kushwaha, Nav Raj Adhikari, Birendra Prasad, Suresh Kumar Maurya, Devarajan Thangadurai and Jeyabalan Sangeetha</i>		6.2.3 bZIP	81
5.1 Introduction	61	6.2.4 DNA binding with one finger family	81
5.1.1 Genome	61	6.2.5 MADS family	82
5.1.2 DNA sequencing	62	6.2.6 Myeloblastosis family	82
5.1.3 Research areas	63	6.2.7 NAM/ATAF/CUC family	82
5.1.4 Model systems for the study of genome	64	6.2.8 WRKY family	83
5.2 Development of genomic resources	64	6.2.9 Zinc fingers	83
5.2.1 Molecular markers	64	6.3 Transcription factors for biotic and abiotic tolerance	84
5.2.2 Transcriptome assemblies	65	6.4 Transcription factor databases	84
5.2.3 Biparental mapping populations	65	6.5 Bioinformatics tools used for structural and functional analysis of transcription factor gene families	84
5.2.4 Genetic linkage maps	66	6.5.1 Data mining by National Center for Biotechnology Information	94
5.2.5 Comparative genome mapping	66	6.5.2 BLAST tool	97
5.2.6 Functional genomics	66	6.5.3 Multiple sequence alignment	97
5.3 Application of genomic resources for crop improvement	66	6.5.4 Physicochemical properties analysis	98
5.3.1 Genetic fingerprinting	66	6.5.5 Motif and domain prediction	98
5.3.2 Hybrid testing	66	6.5.6 In silico structure prediction of proteins	99
5.3.3 Marker-assisted selection	68	6.5.7 Gene predictions	100

6.5.8 Gene duplication and functional divergence studies	100	8.4.1 Metabolomics for biotic and abiotic stresses assessment	126
6.6 Conclusion	100	8.4.2 Metabolomics for soils science and soil conservation	127
References	100	8.4.3 Metabolomics for crops production	129
7. Proteomics as a tool to understand the biology of agricultural crops	107	8.4.4 Metabolomics for crops quality	129
<i>Riyazuddin Riyazuddin,</i>		8.4.5 Metabolomics and postharvest crops science	130
<i>Ashish Kumar Choudhary, Nisha Khatri,</i>		8.5 Conclusions and future perspectives	132
<i>Abhijit Sarkar, Ganesh Kumar Agrawal,</i>		References	132
<i>Sun Tae Kim, Ravi Gupta and Randeep Rakwal</i>		9. Plant metabolomics: a new era in the advancement of agricultural research	139
7.1 Introduction	107	<i>Priyanka Narad, Romasha Gupta and Abhishek Sengupta</i>	
7.2 Gel-based proteomics	108	9.1 An introduction to metabolomics	139
7.2.1 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis	110	9.2 Significance of metabolomics in plant biotechnology	140
7.2.2 Two-dimensional gel electrophoresis	110	9.3 Technologies involved in metabolomics improvement	141
7.2.3 Two-dimensional-difference-in-gel electrophoresis	110	9.4 Metabolomics databases	142
7.3 Gel-free proteomics	111	9.5 Metabolite profiling, identification, and quantification	144
7.3.1 Multidimensional Protein Identification Technology	111	9.6 Metabolic engineering in plants	144
7.3.2 Sequential window acquisition of all theoretical mass spectra	111	9.7 Environmental and ecological metabolomics	146
7.3.3 Label-free quantification	112	9.8 Extraction methods in metabolomics	147
7.3.4 Isobaric tags for relative and absolute quantitation	112	9.9 Metabolomics-assisted breeding techniques	148
7.3.5 Tandem mass tag	112	9.9.1 Metabolic quantitative trait loci	149
7.3.6 Stable Isotope Labeling by Amino acids in Cell Culture	113	9.9.2 Metabolic genome-wide association studies	149
7.4 High-throughput posttranslational modification proteomics	113	9.10 Metabolites present in plant metabolome	150
7.4.1 Phosphorylation	114	9.11 Workflow of metabolomics analysis	151
7.4.2 Glycosylation	114	9.11.1 Sample preparation	151
7.4.3 Acetylation	115	9.11.2 Data mining, annotation, and processing in metabolomics	152
7.5 Conclusion	116	9.11.3 Statistical tools and biomarker identification	152
References	116	9.12 Current and emerging methodologies of metabolomics in agriculture	153
Further reading	122	9.13 Integration of metabolomics tools with other omics tools	154
8. Metabolomics and sustainable agriculture: concepts, applications, and perspectives	123	9.14 Metabolomics under normal and stress conditions in plants	155
<i>Noureddine Benkeblia</i>		9.14.1 Drought stress	155
8.1 Introduction	123	9.14.2 Salinity stress	155
8.2 Sustainable agriculture and agro-production systems	124	9.14.3 Waterlogging stress	155
8.3 Concepts of metabolomics and their applications to agriculture	124	9.14.4 Temperature stress	156
8.4 Bridging metabolomics to sustainable agriculture	126	9.14.5 Metal-induced stress	156

9.15 Applications and future perspective of metabolomics in plant biotechnology and agriculture	156	11.3 Long noncoding RNA	181
References	157	11.4 Circular RNA	184
		11.5 Chimeric RNA	185
		References	186
10. Explore the RNA-sequencing and the next-generation sequencing in crops responding to abiotic stress	161	12. Molecular evolution, three-dimensional structural characteristics, mechanism of action, and functions of plant beta-galactosidases	191
<i>Éderson Akio Kido,</i>		<i>Md. Anowar Hossain</i>	
<i>José Ribamar Costa Ferreira-Neto,</i>		12.1 Introduction	191
<i>Eliseu Binneck, Manassés da Silva,</i>		12.2 Protein sequence features of plant beta-galactosidases	192
<i>Wilson da Silva Júnior</i>		12.3 Molecular evolution of beta-galactosidases and their classification	198
<i>and Ana Maria Benko-Iseppon</i>		12.4 Three-dimensional structural characteristics of plant beta-galactosidases	200
10.1 Introduction	161	12.5 Structural comparison between MiBGAL and TBG4	201
10.2 From the beginning to the crop sciences: transcriptome analysis, its evolution, and state of the art	161	12.6 Substrate specificity of plant beta-galactosidases	201
10.3 The overview on plant sequencing of RNA studies	163	12.7 Mechanism of action of plant beta-galactosidases	204
10.4 The RNA-sequencing analysis workflow	165	12.8 Physiological function of plant beta-galactosidase	205
10.4.1 Data generation	165	12.9 Conclusion	206
10.4.2 Raw data processing	168	Conflict of interest	206
10.4.3 Data analysis	169	References	206
10.4.4 Accessing the overall quality of the assembly	169		
10.4.5 Transcript quantification	170	13. Next generation genomics: toward decoding domestication history of crops	209
10.4.6 Differential expression analysis	170	<i>Anjan Hazra and Sauren Das</i>	
10.4.7 Annotation and functional analysis	170	13.1 Introduction	209
10.5 Functional genomics	171	13.2 Whole genome sequencing	209
10.6 Final considerations	171	13.3 Alternative genome scale approaches	210
Acknowledgments	172	13.4 Emergence of pan-genomics	212
References	172	13.5 Methodologies in domestication genomics	212
		13.6 Case studies on next-generation sequencing-assisted inference of domestication history	212
11. Identification of novel RNAs in plants with the help of next-generation sequencing technologies	177	13.6.1 Rice	212
<i>Aditya Narayan and Shailesh Kumar</i>		13.6.2 Citrus	215
11.1 Introduction	177	13.6.3 Peanut	215
11.1.1 Noncoding RNA classes in plants	177	13.6.4 Olive	215
11.2 Small RNA	177	13.6.5 Tea	216
11.2.1 MicroRNA	177	References	216
11.2.2 Small-interfering RNA	179		
11.2.3 Heterochromatic small-interfering RNA	180		
11.2.4 Phased small-interfering RNA and trans-acting small-interfering RNA	180		
11.2.5 Natural antisense-small-interfering RNA	180		
11.2.6 Transfer RNA–derived small RNA	181		

14. <i>In-silico</i> identification of small RNAs: a tiny silent tool against agriculture pest	221	15.6 Conclusion	245
<i>Habeeb Shaik Mohideen, Kevina Sonawala and Sewali Ghosh</i>		Acknowledgments	245
14.1 Introduction	221	References	245
14.2 Small RNAs	221	16. Omics-assisted understanding of BPH resistance in rice: current updates and future prospective	253
14.3 Types of small noncoding RNAs	222	<i>Satyabrata Nanda</i>	
14.4 Next-generation sequencing in agronomic advancements	222	16.1 Introduction	253
14.5 Small RNA world and their identification	222	16.2 Rice genomics in brown planthopper resistance	253
14.5.1 MicroRNA	222	16.3 Rice transcriptomics in brown planthopper resistance	256
14.5.2 PIWI-interacting RNAs	225	16.4 Rice proteomics in brown planthopper resistance	257
14.5.3 Small interfering RNAs	226	16.5 Rice metabolomics in brown planthopper resistance	258
14.6 Limitations	226	16.6 Bioinformatics in brown planthopper resistance in rice	259
14.7 Conclusion	227	16.7 Conclusion and future prospective	259
Acknowledgments	228	References	259
References	228	17. Contemporary genomic approaches in modern agriculture for improving tomato varieties	265
Section II		<i>Nikolay Manchev Petrov, Mariya Ivanova Stoyanova, Rajarshi Kumar Gaur, Milena Georgieva Bozhilova-Sakova and Ivona Vassileva Dimitrova</i>	
Omics application		17.1 Importance and origin of tomatoes	265
15. Bioinformatics-assisted multiomics approaches to improve the agronomic traits in cotton	233	17.2 Organization of tomato genome and genetic variation of tomato cultivars	267
<i>Sidra Aslam, Muhammad Aamer Mehmood, Mehboob-ur Rahman, Fatima Noor and Niaz Ahmad</i>		17.3 Tomato breeding	269
15.1 Introduction	233	17.4 Disease resistance	269
15.1.1 A bird's-eye view of the world cotton market	233	17.5 Insect resistance	270
15.1.2 An overview of omics mainly focused on plant-omics	233	17.6 Abiotic stress tolerance	270
15.1.3 Introduction of bioinformatics in the area of next-generation sequencing	234	17.7 Tomato genetic markers for selection	271
15.1.4 Brief description of "integration of omics"	235	17.8 Genomic selection for abiotic stress in tomato	271
15.1.5 Why is multiomics study preferred over single-omics?	236	17.9 Tomato transcriptomics	272
15.2 Big data in biology and omics	237	17.10 Tomato proteomics	272
15.3 Bioinformatics resources for cotton-omics	237	17.11 Tomato metabolomics	272
15.3.1 Genomics	240	References	272
15.3.2 Proteomics	242	18. Characterization of drought tolerance in maize: omics approaches	279
15.3.3 Metabolomics	243	<i>Ramandeep Kaur, Manjot Kaur, Parampreet Kaur and Priti Sharma</i>	
15.4 Integration of multiomics data to cope with cotton plant diseases	243	18.1 Introduction	279
15.5 Challenges in the integration and analysis of multiomics data of cotton	244		

18.2	Drought timing	280	20.2	Overview of molecular marker systems in wheat	324
18.3	Plant response to drought	281	20.3	Genome-wide markers for gene mapping	328
18.4	Progress with conventional breeding strategies for drought tolerance in maize	282	20.4	Wheat genomics for development of marker and its utilization	328
18.4.1	Seedling and physiological traits for drought tolerance	283	20.5	Status of genotyping platform of bread wheat and its progenitors	329
18.4.2	Yield traits for drought tolerance	283	20.5.1	High-throughput SNP genotyping: microarray-based genotyping	329
18.5	Omics for characterizing drought stress responses in maize	284	20.5.2	High-throughput SNP genotyping: genotyping-by-sequencing	330
18.5.1	Genomics	284	20.6	Utility and achievement of high-throughput genotyping approaches in wheat	331
18.5.2	Transcriptomics	287	20.7	Conversion of trait-linked SNPs to user-friendly markers	332
18.5.3	Proteomics and metabolomics	288	20.8	Conclusions and future directions	333
18.5.4	Advances in phenomics	289	References	333	
18.5.5	Bioinformatics tools and databases	289			
18.6	Conclusion	290			
References		290			
19.	Deciphering the genomic hotspots in wheat for key breeding traits using comparative and structural genomics	295	21.	Omics approaches for biotic, abiotic, and quality traits improvement in potato (<i>Solanum tuberosum</i> L.)	341
	<i>Dharmendra Singh, Pritesh Vyas, Chandranandani Negi, Imran Sheikh and Kunal Mukhopadhyay</i>			<i>Jagesh Kumar Tiwari, Tanuja Buckseth, Clarissa Challam, Nandakumar Natarajan, Rajesh K. Singh and Manoj Kumar</i>	
19.1	Introduction	295	21.1	Introduction	341
19.2	Genomic comparisons and gene discovery	296	21.2	Potato genomics	342
19.2.1	Gene discovery and marker development	296	21.2.1	Whole-genome sequencing and resequencing	342
19.2.2	Gene annotation and marker development	297	21.2.2	Molecular markers	342
19.2.3	Functional comparative genomics in cereals	298	21.2.3	Quantitative trait loci mapping, bulked segregant analysis, and GWAS	343
19.3	Genomic hotspots in wheat	298	21.3	Potato transcriptomic	344
19.3.1	Biofortification hotspots	298	21.3.1	Biotic stress	346
19.3.2	Genomic hotspots for biotic stress resistance	300	21.3.2	Abiotic stress	347
19.3.3	Genomic hotspots for drought stress tolerance	303	21.3.3	Quality traits	347
19.3.4	Genomic hotspots for heat tolerance in wheat	304	21.3.4	miRNAs in potato	348
19.4	Genomic sequences to genomic hotspot	305	21.4	Potato proteomics	348
19.5	Conclusion	314	21.4.1	Biotic stress	348
References		314	21.4.2	Abiotic stress	350
			21.4.3	Quality traits	350
20.	Prospects of molecular markers for wheat improvement in postgenomic era	323	21.5	Potato metabolomics	350
	<i>Satish Kumar, Disha Kamboj, Chandra Nath Mishra and Gyanendra Pratap Singh</i>		21.5.1	Biotic traits	351
20.1	Introduction	323	21.5.2	Abiotic traits	351
			21.5.3	Quality traits	351
			21.6	Potato ionomics	352
			21.7	Phenomics	352
			21.8	Potato omics resources and integration of technologies	353
			21.9	Conclusions	354
			References	354	

22. Tea plant genome sequencing: prospect for crop improvement using genomics tools	361	24.5 Challenges	392
<i>Pradosh Mahadani and Basant K. Tiwary</i>		24.6 Conclusion and future prospective	393
22.1 Introduction	361	References	393
22.2 Whole-genome sequencing of tea plant	363	25. Microbial degradation of herbicides in contaminated soils by following computational approaches	399
22.3 Identification and characterization of gene families	365	<i>Kusum Dhakar, Hanan Eizenberg, Zeev Ronen, Raphy Zarecki and Shiri Freilich</i>	
22.4 Tea transcriptome sequencing	365	25.1 Herbicides: use and impact on environment	399
22.5 Discovery of single-nucleotide polymorphism	368	25.2 Microbial degradation of herbicides	400
22.6 Conclusion	369	25.3 Strategies to improve biodegradation of herbicides	402
References	369	25.4 Integration of computational biology to improve biodegradation of herbicides	406
23. Next-generation sequencing and viroid research	373	25.5 Bioremediation of atrazine by following metabolic modeling method	409
<i>Sunny Dhir, Asha Rani and Narayan Rishi</i>		25.6 Conclusion	410
23.1 Introduction	373	Acknowledgments	410
23.2 Next-generation sequencing technology	374	References	410
23.3 Impact of next-generation sequencing on viroid discovery	375	26. Chloroplast genome and plant–virus interaction	419
23.4 Role of next-generation sequencing in unraveling viroid RNA biology	376	<i>Parampreet Kaur, Tanvi Kaila, Manmohan Dhkal and Kishor Gaikwad</i>	
23.4.1 Characterization of viroid sequence variants	376	26.1 Introduction	419
23.4.2 Viroid pathogenesis	376	26.2 Chloroplast genome	420
23.4.3 Mutational analyses of the viroids	379	26.2.1 Structure and gene content	420
23.5 Bioinformatic intervention in next-generation sequencing	379	26.2.2 Genomic advances	422
23.6 Conclusion	380	26.2.3 Bioinformatic approaches and plastomes	422
References	380	26.2.4 Status of chloroplast genome sequencing in plants	423
24. Computational analysis for plant virus analysis using next-generation sequencing	383	26.3 Viral infection symptoms in plants	423
<i>Chitra Nehra, Rakesh Kumar Verma, Nikolay Manchev Petrov, Mariya Ivanova Stoyanova, Pradeep Sharma and Rajarshi Kumar Gaur</i>		26.4 Role of chloroplasts in plant–virus life cycle	426
24.1 Introduction	383	26.4.1 Changes in chloroplast structure upon viral infection	427
24.2 Development of next-generation sequencing technology	383	26.4.2 Virus factors involved in structural and functional changes of chloroplast	427
24.3 Next-generation sequencing data analysis by bioinformatics tools	385	26.5 Role of chloroplast in the defense against plant pathogenic viruses	428
24.4 Next-generation sequencing in plant virology	386	26.6 Plant–virus metagenomics	429
		26.7 Conclusion	430
		References	430

Section III

Data mining, markers discovery**27. Deciphering soil microbiota using metagenomic approach for sustainable agriculture: an overview 439***Aiman Tanveer, Shruti Dwivedi, Supriya Gupta, Rajarshi Kumar Gaur and Dinesh Yadav*

27.1 Introduction	439
27.2 Sustainable agriculture	439
27.3 Soil microbiomes	440
27.4 Soil microbial diversity	441
27.5 Analysis of the rhizosphere microbial community	441
27.6 Metagenomics in agriculture	442
27.6.1 Metagenomics based techniques for rhizosphere analysis	443
27.7 Metagenomics for sustainable agriculture	446
27.8 Concluding remarks	449
References	450

28. Concepts and applications of bioinformatics for sustainable agriculture 455*Ezgi Çabuk Şahin, Yıldız Aydın, Tijs Gilles, Ahu Altinkut Uncuoğlu and Stuart J. Lucas*

28.1 Introduction—a conceptual framework for sustainable agriculture	455
28.2 Database resources for agricultural bioinformatics	455
28.3 Genome mapping	459
28.3.1 Molecular marker systems and populations used for genetic mapping	459
28.3.2 Genetic mapping, physical mapping, and genome sequencing	461
28.3.3 Comparative mapping	461
28.3.4 Practical applications of genetic mapping	462
28.4 DNA marker development and application to genotyping	462
28.4.1 DNA marker types, their advantages and disadvantages	463
28.4.2 Shift to single-nucleotide polymorphism and insertion/deletion markers	465
28.4.3 Genotyping technologies and their application in breeding programs	465

28.4.4 Medium-throughput genotyping technologies	465
28.4.5 High-throughput genotyping technologies	467
28.4.6 Increased automation and throughput while reducing cost per data point	469
28.4.7 Single-nucleotide polymorphism genotyping for sustainable agriculture in a complex genome—bread wheat	469

28.5 Genome-wide association studies 474

28.5.1 Using single-nucleotide polymorphism markers for genome-wide association studies	474
28.5.2 Genome-wide association studies' design and analysis	475
28.5.3 Applications of genome-wide association studies to plant and animal breeding	476

28.6 Emerging strategies for breeding and genetics 477

28.6.1 Gene expression regulation by noncoding RNA	477
28.6.2 Translation of “omics” data to agriculture	479
28.6.3 Bioinformatic resources for sustainable crop and livestock production	479

28.7 Conclusion and future prospects 480

References	481
------------	-----

29. Application of high-throughput structural and functional genomic technologies in crop nutrition research 491*Nand Lal Meena, Ragini Bhardwaj, Om Prakash Gupta, Vijay Singh Meena, Ajeet Singh and Aruna Tyagi*

29.1 Introduction	491
29.2 Structural genomics	491
29.3 Application of structural genomics	492
29.3.1 To determine each single protein structure encrypted by the genome	492
29.3.2 Identification of three-dimensional structure and folding of novel protein functions	493
29.3.3 Gene and protein interactions: the role of protein structure prediction in structural genomics	493

29.4	Dynamic expression of functional genomics	493	31.4	Single-nucleotide polymorphism database	520
29.5	Functional genomics approaches	494	31.5	Single-nucleotide polymorphism genotyping	521
29.6	Developing genomic technologies for enhancing food crops security	496	31.5.1	Gel-based single-nucleotide polymorphism genotyping	522
29.7	Application of high-throughput genomics technologies in nutrition research	496	31.5.2	Nongel-based single-nucleotide polymorphism genotyping	523
	References	497	31.6	Application of single-nucleotide polymorphisms in plants	524
	Further reading	498	31.6.1	Genetic diversity	525
30.	Bioinformatics approach for whole transcriptomics-based marker prediction in agricultural crops	503	31.6.2	Genetic mapping	525
	<i>Habeeb Shaik Mohideen, Archit Gupta and Sewali Ghosh</i>		31.6.3	Phylogenetic analysis	526
30.1	Introduction to transcriptomics	503	31.6.4	Marker-assisted selection	526
30.1.1	Transcriptome	503	31.7	Conclusion and prospects	528
30.2	Markers	503		Acknowledgment	528
30.2.1	Phenotypic markers	503		References	528
30.2.2	Biochemical markers	504	32.	Bioinformatics intervention in identification and development of molecular markers: an overview	537
30.2.3	Cytological markers	504		<i>Vikas Dwivedi, Lalita Pal and Dinesh Yadav</i>	
30.2.4	Molecular markers	504	32.1	Introduction	537
30.3	Markers in plants	504	32.2	Genetic markers	538
30.4	Expressed sequence tags and simple sequence repeats	505	32.2.1	Classical markers: The classical markers are further divided that include morphological markers, cytological markers and biochemical markers	538
30.5	Tools for generating transcriptomic data	505	32.2.2	Molecular markers	538
30.5.1	Serial analysis of gene expression technology	505	32.3	Restriction fragment length polymorphism (RFLP)	538
30.5.2	Microarrays	505	32.3.1	Application of restriction fragment length polymorphism	539
30.5.3	RNA sequencing	505	32.4	Random amplified polymorphic DNA (RAPD)	541
30.6	Why transcriptomic markers?	506	32.4.1	Applications of random amplified polymorphic DNA	541
30.7	How are markers developed/selected?	507	32.5	Amplified fragment length polymorphism (AFLP)	542
30.8	What has been done	508	32.5.1	Advantages of amplified fragment length polymorphism	542
30.9	Future prospects	508	32.5.2	Disadvantages of amplified fragment length polymorphism	543
	References	509	32.5.3	Techniques for amplified fragment length polymorphism data analysis	544
31.	Computational approaches toward single-nucleotide polymorphism discovery and its applications in plant breeding	513	32.5.4	Application of amplified fragment length polymorphism	544
	<i>Dileep Kumar, Ranjana Gautam, Veda P. Pandey, Anurag Yadav, Upendra N. Dwivedi, Rumana Ahmad and Kusum Yadav</i>		32.6	Simple sequence repeats (SSR)	544
31.1	Introduction	513	32.6.1	Distribution of simple sequence repeats	545
31.2	Single-nucleotide polymorphism discovery	514			
31.2.1	Reference-based single-nucleotide polymorphism mining	514			
31.2.2	De novo single-nucleotide polymorphism discovery	517			
31.3	Single-nucleotide polymorphism annotation	518			

32.6.2	Isolation of simple sequence repeats markers	545		
32.6.3	Applications of microsatellite	546		
32.7	Intersimple sequence repeat (ISSR)	546		
32.7.1	Advantages of intersimple sequence repeat markers	546		
32.7.2	Disadvantages of intersimple sequence repeat markers	547		
32.7.3	Application of intersimple sequence repeat markers	547		
32.8	Single-nucleotide polymorphism (SNP)	547		
32.8.1	Single-nucleotide polymorphism detection	547		
32.8.2	In vitro techniques	547		
32.8.3	Single-nucleotide polymorphism application	548		
32.8.4	Diversity array technology (DArT Seq)	548		
32.9	Quantitative trait loci (QTL)	548		
32.9.1	Molecular markers	548		
32.9.2	Construction of genetic linkage maps	548		
32.9.3	Mapping population	549		
32.9.4	Identification of polymorphism	549		
32.9.5	Linkage analysis of markers	549		
32.9.6	Genetic distance and mapping functions	550		
32.9.7	Quantitative trait loci analysis	550		
32.9.8	Quantitative trait loci detection	550		
32.9.9	Advantages and disadvantages of quantitative trait loci mapping	550		
32.10	Association mapping	551		
32.10.1	Linkage disequilibrium	551		
32.10.2	Methods of association mapping	551		
32.10.3	Class of association mapping	551		
32.10.4	Association mapping in the breeding program	552		
32.11	Marker-assisted selection (MAS)	552		
32.11.1	Application of marker-assisted selection	552		
32.12	Bioinformatics intervention in molecular markers	553		
32.13	Software for simple sequence repeats discovery	553		
32.14	Software for single-nucleotide polymorphism discovery	555		
References		555		
33.	Deciphering comparative and structural variation that regulates abiotic stress response			561
	<i>Zeba Seraj, Sabrina Elias, Saima Shahid, Taslima Haque, Richard Malo and Mohammad Umer Sharif Shohan</i>			
33.1	Introduction			561
33.2	Expression quantitative trait loci and their functional significance			562
33.2.1	Molecular marker system for genotyping			562
33.2.2	Transcript abundance measurement by RNA sequence			564
33.2.3	Connecting genomic variation to expression variation			565
33.3	Regulatory small RNAs			566
33.3.1	Discovery and annotation of small RNAs based on deep sequencing			566
33.3.2	Detection of small RNA targets			568
33.3.3	Natural variation in small RNAs and their targets			568
33.3.4	Integrating small RNA sequencing with quantitative trait loci mapping			568
33.4	Epigenomic regulation of gene expression in plant			569
33.4.1	DNA methylation and its role in transcriptional regulation			569
33.4.2	The role of histone modification for the regulation of gene expression			571
33.5	Protein structure provides vital information of function during salt stress			572
33.5.1	Variation in protein structure contributing to salinity tolerance			572
33.5.2	Future prospect in substitution-mediated enhanced salt tolerance			574
33.6	High performance computing in comparative genomics			574
33.7	Conclusion			579
References				580

Section IV Artificial intelligence and agribots

34. Deep Learning applied to computational biology and agricultural sciences 589

Renato Hidaka Torres, Fabricio Almeida Araujo, Edian Franklin Franco De Los Santos, Debmalya Barh, Rommel Thiago Jucá Ramos and Marcus de Barros Braga

34.1 Introduction	589
34.2 Deep Learning and Convolutional Neural Network	589
34.3 Deep Learning applications in computational biology	593
34.3.1 Omics	593
34.3.2 Biological image processing	595
34.3.3 Multiomic data integration	596
34.3.4 Single-cell RNA sequencing	597
34.3.5 Pharmacogenomics	597
34.3.6 Modeling biological data in a Deep Neural Network	598
34.4 Deep Learning applications in agricultural sciences	600
34.4.1 Example of Deep Learning applied to agriculture	601
34.4.2 Convolutional Neural Networks in agriculture	601
34.4.3 Recurrent Neural Network for agricultural classification	608
34.5 Conclusion	612
References	612

35. Image processing–based artificial intelligence system for rapid detection of plant diseases 619

Sanjaya Shankar Tripathy, Raju Poddar, Lopamudra Satapathy and Kunal Mukhopadhyay

35.1 Introduction	619
35.2 Visual symptoms of diseases in plant	619
35.3 Imaging	620
35.4 Database creation	621
35.5 Disease identification using feature extraction and classification	621
35.6 Disease identification using convolutional neural network	622
35.7 Determination of the accuracy of the system	623
35.8 Severity estimation	623
35.9 Conclusion	624
References	624

36. Role of artificial intelligence, sensor technology, big data in agriculture: next-generation farming 625

Pradeep Kumar, Abhishek Singh, Vishnu D. Rajput, Ajit Kumar Singh Yadav, Pravin Kumar, Anil Kumar Singh and Tatiana Minkina

36.1 Introduction	625
36.2 Characteristics of big data	626
36.2.1 Volume	626
36.2.2 Velocity	626
36.2.3 Variety	627
36.2.4 Veracity	627
36.3 Big data and smart agriculture	627
36.3.1 Digital soil and crop mapping	627
36.3.2 Weather prediction	627
36.3.3 Fertilizers recommendation	628
36.3.4 Disease detection and pest management	628
36.3.5 Adaptation to climate change	628
36.3.6 Automated irrigation system	628
36.4 Sources of big data	628
36.4.1 Sensors	629
36.4.2 Statistical data	630
36.4.3 Remote sensing	631
36.4.4 Cloud data source	631
36.4.5 Internet of things database source	631
36.4.6 Media source	631
36.5 Techniques and tool use in big data analysis	631
36.5.1 Machine learning	632
36.5.2 Cloud platforms	633
36.5.3 Geographic information systems	633
36.5.4 Vegetation indices	633
36.6 Role of big data in agriculture ecosystem: for smart farming	635
Acknowledgments	636
Conflict of interest	636
Author contributions	637
References	637

37. Artificial intelligence: a way forward for agricultural sciences 641

Neeru S. Redhu, Zoozeal Thakur, Shikha Yashveer and Poonam Mor

37.1 Introduction of artificial intelligence	641
37.2 History of artificial intelligence	642
37.3 Methods and approaches in artificial intelligence	643
37.3.1 Machine learning	645
37.3.2 Artificial neural network	649
37.3.3 Deep learning	649

37.4 Technological advancements in artificial intelligence	650	37.5.3 Biological sciences	662
37.4.1 Hardware	651	37.6 Future perspective and challenges	663
37.4.2 Software	655	37.7 Conclusion	664
37.5 Application of artificial intelligence	656	References	664
37.5.1 Agriculture/farming	658	Index	669
37.5.2 As a service industry	661		

List of contributors

- Nav Raj Adhikari** Institute of Agriculture and Animal Science, Tribhuvan University, Kirtipur, Kathmandu, Nepal
- Ganesh Kumar Agrawal** Research Laboratory for Biotechnology and Biochemistry (RLABB), Kathmandu, Nepal
- Niaz Ahmad** Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering, Faisalabad, Pakistan; Department of Biotechnology, Pakistan Institute of Engineering & Applied Sciences (PIEAS), Islamabad, Pakistan
- Rumana Ahmad** Department of Biochemistry, Era University, Lucknow, Uttar Pradesh, India
- Fabricio Almeida Araujo** Federal University of Pará, Belém, Brazil
- Sidra Aslam** Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Faisalabad, Pakistan
- Yıldız Aydın** Department of Biology, Faculty of Science & Arts, Marmara University, Istanbul, Turkey
- Debmalya Barh** Institute of Integrative Omics and Applied Biotechnology, Purba Medinipur, West Bengal, India
- Noureddine Benkeblia** Department of Life Sciences, The Biotechnology Center, The University of the West Indies, Kingston, Jamaica
- Ana Maria Benko-Iseppon** Federal University of Pernambuco, University in Recife, Brazil
- Ragini Bhardwaj** ICAR-National Bureau of Plant Genetic Resources, New Delhi, Delhi, India
- Eliseu Binneck** Brazilian Agricultural Research Corporation, University in Recife, Brazil
- Milena Georgieva Bozhilova-Sakova** Department “Genetics, Breeding, Selection, Reproduction and Biotechnologies of Farm Animals”, Institute of Animal Science, Kostinbrod, Bulgaria
- Marcus de Barros Braga** Federal Rural University of Amazonia, Paragominas, Brazil
- Tanuja Buckseth** ICAR-Central Potato Research Institute, Shimla, Himachal Pradesh, India
- Clarissa Challam** ICAR-Central Potato Research Institute, Regional Station, Shillong, Meghalaya, India
- Ashish Kumar Choudhary** Department of Botany, University of Delhi, Delhi, India
- Manassés da Silva** Federal University of Pernambuco, University in Recife, Brazil
- Wilson da Silva, Júnior** Federal University of Pernambuco, University in Recife, Brazil
- Sauren Das** Agricultural and Ecological Research Unit, Indian Statistical Institute, Kolkata, India
- Edian Franklin Franco De Los Santos** Federal University of Pará, Belém, Brazil; Santo Domingo Technological Institute, Santo Domingo, Dominican Republic
- Kusum Dhakar** Newe Ya’ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel; Department of Environmental Hydrology & Microbiology, Zuckerberg Institute for Water Research, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel
- Sunny Dhir** Amity Institute of Virology & Immunology, Amity University, Noida, Uttar Pradesh, India
- Manmohan Dhal** School of Organic Farming, Punjab Agricultural University, Ludhiana, India
- Ivona Vassileva Dimitrova** Department of Plant Protection, Agronomy Faculty, University of Forestry, Sofia, Bulgaria
- Shruti Dwivedi** Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India
- Upendra N. Dwivedi** Department of Biochemistry, University of Lucknow, Lucknow, Uttar Pradesh, India; Institute for Development of Advanced Computing, ONGC Center for Advanced Studies, University of Lucknow, Lucknow, Uttar Pradesh, India

- Vikas Dwivedi** Agricultural Research Organization, The Volcani Center, Rishon LeZion, Israel; National Institute of Plant Genome Research, New Delhi, New Delhi, India
- Hanan Eizenberg** Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel
- Sabrina Elias** Department of Life Sciences, Independent University Bangladesh, Dhaka, Bangladesh
- José Ribamar Costa Ferreira-Neto** Federal University of Pernambuco, University in Recife, Brazil
- Shiri Freilich** Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel
- Kishor Gaikwad** ICAR-National Institute for Plant Biotechnology, New Delhi, India
- Rajarshi Kumar Gaur** Department of Biotechnology, Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India
- Ranjana Gautam** Department of Biochemistry, University of Lucknow, Lucknow, Uttar Pradesh, India
- Sewali Ghosh** Department of Zoology and Advanced Biotechnology, Guru Nanak College, Chennai, Tamil Nadu, India
- Tijs Gilles** Sabanci University Nanotechnology Research and Application Center, Istanbul, Turkey
- Archit Gupta** Bioinformatics and Entomoinformatics Lab, Department of Genetic Engineering, School of Bioengineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India
- Kapil Gupta** Department of Biotechnology, Siddharth University, Siddharthnagar, Uttar Pradesh, India
- Om Prakash Gupta** ICAR-Indian Institute of Wheat and Barley Research, Karnal, Haryana, India
- Ravi Gupta** Department of Botany, School of Chemical and Life Sciences, Jamia Hamdard, New Delhi, India
- Romasha Gupta** Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India
- Supriya Gupta** Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India
- Taslina Haque** Department of Integrative Biology, University of Texas at Austin, Austin, TX, United States
- Anjan Hazra** Agricultural and Ecological Research Unit, Indian Statistical Institute, Kolkata, India
- Md. Anowar Hossain** Department of Biochemistry and Molecular Biology, University of Rajshahi, Rajshahi, Bangladesh
- Niranjani Iyer** Biovia Corp, Dassault Systemes, San Diego, CA, United States
- Akanksha Jaiswar** Centre for Agricultural Bioinformatics (CABin), ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India
- Rahul Singh Jasrotia** Department of Computational Biology & Bioinformatics, JIBB Sam Higginbottom University of Agriculture, Technology & Sciences, (Formerly AAI-DU), Prayagraj (Allahabad), Uttar Pradesh, India
- Tanvi Kaila** ICAR-National Institute for Plant Biotechnology, New Delhi, India
- Disha Kamboj** ICAR-Indian Institute of Wheat and Barley Research Institute, Karnal, Haryana, India
- Manjot Kaur** School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India
- Parampreet Kaur** School of Organic Farming, Punjab Agricultural University, Ludhiana, India
- Ramandeep Kaur** School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India
- Nisha Khatri** Department of Botany, Dyal Singh College, University of Delhi, Delhi, India
- Éderson Akio Kido** Federal University of Pernambuco, University in Recife, Brazil
- Sun Tae Kim** Department of Plant Bioscience, Pusan National University, Miryang, Republic of Korea
- Ashwani Kumar** Department of Agricultural Biotechnology, College of Agriculture, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India
- Dileep Kumar** Department of Biochemistry, University of Lucknow, Lucknow, Uttar Pradesh, India
- Manoj Kumar** ICAR-Central Potato Research Institute, Shimla, Himachal Pradesh, India
- Pradeep Kumar** Department of Forestry, Applied Microbiology Laboratory, North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India
- Pravin Kumar** Department of Electrical Engineering, School of Engineering, Gautam Buddha University, Greater Noida, Uttar Pradesh, India
- Satish Kumar** ICAR-Indian Institute of Wheat and Barley Research Institute, Karnal, Haryana, India
- Shailesh Kumar** Bioinformatics Laboratory, National Institute of Plant Genome Research (NIPGR), Delhi, New Delhi, India

- Ujjawal Kumar Singh Kushwaha** National Plant Breeding and Genetics Research Center, Nepal Agricultural Research Council, Khumaltar, Lalitpur, Nepal
- Stuart J. Lucas** Sabanci University Nanotechnology Research and Application Center, Istanbul, Turkey
- Pradosh Mahadani** National Tea Research Foundation, Tea Board, Kolkata, India
- Richard Malo** Life Science Division, Overseas Marketing Corporation Pvt. Ltd., Dhaka, Bangladesh
- Suresh Kumar Maurya** Department of Vegetable Science, College of Agriculture, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Utrakhnad, India
- Nand Lal Meena** ICAR-National Bureau of Plant Genetic Resources, New Delhi, Delhi, India; Division of Biochemistry, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
- Vijay Singh Meena** ICAR-National Bureau of Plant Genetic Resources, New Delhi, Delhi, India
- Muhammad Aamer Mehmood** Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Faisalabad, Pakistan
- Tatiana Minkina** Academy of Biology and Biotechnology, Southern Federal University, Rostov-on-Don, Russia
- Chandra Nath Mishra** ICAR-Indian Institute of Wheat and Barley Research Institute, Karnal, Haryana, India
- Habeeb Shaik Mohideen** Bioinformatics and Entomoinformatics Lab, Department of Genetic Engineering, School of Bioengineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India
- Poonam Mor** Bacteriology Lab, National Research Centre on Equines, Hisar, Haryana, India
- Kunal Mukhopadhyay** Department of Bioengineering and Biotechnology, Birla Institute of Technology, Mesra, Jharkhand, India
- Satyabrata Nanda** MS Swaminathan School of Agriculture, Centurion University of Technology and Management, Paralakhemundi, Odisha, India
- Priyanka Narad** Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India
- Aditya Narayan** University of Virginia, Charlottesville, VA, United States
- Nandakumar Natarajan** ICAR-Central Potato Research Institute, Regional Station, Shillong, Meghalaya, India
- Chandranandani Negi** Dr. Khem Singh Gill Akal College of Agriculture, Eternal University, Baru Sahib, Himachal Pradesh, India
- Chitra Nehra** Department of Biosciences, Mody University of Science and Technology, Sikar, Rajasthan, India
- Fatima Noor** Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Faisalabad, Pakistan
- Lalita Pal** National Institute of Plant Genome Research, New Delhi, New Delhi, India
- Veda P. Pandey** Department of Biochemistry, University of Lucknow, Lucknow, Uttar Pradesh, India
- Nikolay Manchev Petrov** Department of Natural Sciences, New Bulgarian University, Sofia, Bulgaria
- Raju Poddar** Department of Bioengineering and Biotechnology, Birla Institute of Technology, Mesra, Jharkhand, India
- Birendra Prasad** Department of Genetics and Plant Breeding, College of Agriculture, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Utrakhnad, India
- Mehboob-ur Rahman** Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering, Faisalabad, Pakistan; Department of Biotechnology, Pakistan Institute of Engineering & Applied Sciences (PIEAS), Islamabad, Pakistan
- Vishnu D. Rajput** Academy of Biology and Biotechnology, Southern Federal University, Rostov-on-Don, Russia
- Randeep Rakwal** Research Laboratory for Biotechnology and Biochemistry (RLABB), Kathmandu, Nepal; GRADE (Global Research Arch for Developing Education) Academy Private Limited, Birgunj, Nepal; Faculty of Health and Sport Sciences, University of Tsukuba, Tsukuba, Japan
- Rommel Thiago Jucá Ramos** Federal University of Pará, Belém, Brazil
- Asha Rani** Amity Institute of Virology & Immunology, Amity University, Noida, Uttar Pradesh, India
- Neeru S. Redhu** Department of Molecular Biology, Biotechnology & Bioinformatics, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India
- Narayan Rishi** Amity Institute of Virology & Immunology, Amity University, Noida, Uttar Pradesh, India

- Riyazuddin Riyazuddin** Department of Plant Biology, Faculty of Science and Informatics, University of Szeged, Szeged, Hungary
- Zeev Ronen** Department of Environmental Hydrology & Microbiology, Zuckerberg Institute for Water Research, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel
- Jeyabalan Sangeetha** Department of Environmental Science, Central University of Kerala, Periyar, Kerala, India
- Abhijit Sarkar** Department of Botany, University of Gour Banga, Malda, West Bengal, India
- Lopamudra Satapathy** Faculty of Agriculture, Usha Martin University, Ranchi, Jharkhand, India
- Abhishek Sengupta** Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India
- Zeba Seraj** Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka, Bangladesh
- Saima Shahid** Donald Danforth Plant Science Center, St. Louis, MO, United States
- Pradeep Sharma** ICAR- Indian Institute of Wheat & Barley Research, Karnal, Haryana, India
- Priti Sharma** School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India
- Swati Sharma** Department of Agricultural Biotechnology, College of Agriculture, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India
- Imran Sheikh** Dr. Khem Singh Gill Akal College of Agriculture, Eternal University, Baru Sahib, Himachal Pradesh, India
- Mohammad Umer Sharif Shohan** Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka, Bangladesh
- Abhishek Singh** Department of Agricultural Biotechnology, College of Agriculture, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India
- Ajeet Singh** Division of Biochemistry, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
- Anil Kumar Singh** DeHaat, Agrevolution Pvt. Limited, Gurugram, Haryana, India
- Anuradha Singh** ICAR-National Institute for Plant Biotechnology, New Delhi, India
- Dharmendra Singh** Government Model College, Jhabua, Madhya Pradesh, India
- Gyanendra Pratap Singh** ICAR-Indian Institute of Wheat and Barley Research Institute, Karnal, Haryana, India
- Nisha Singh** ICAR-National Institute for Plant Biotechnology, New Delhi, India
- Rajesh K. Singh** ICAR-Central Potato Research Institute, Shimla, Himachal Pradesh, India
- Kevina Sonawala** Bioinformatics and Entomoinformatics Lab, Department of Genetic Engineering, School of Bioengineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India
- Sudhanshu Srivastava** Department of Biotechnology, D. D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India
- Mariya Ivanova Stoyanova** Department of Plant Protection, Institute of Soil Science, Agrotechnologies and Plant Protection (ISSAPP) “Nikola Pushkarov”, Sofia, Bulgaria
- Aiman Tanveer** Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India
- Zoozeal Thakur** Bacteriology Lab, National Research Centre on Equines, Hisar, Haryana, India
- Devarajan Thangadurai** Department of Botany, Karnatak University, Dharwad, Karnataka, India
- Jagesh Kumar Tiwari** ICAR-Central Potato Research Institute, Shimla, Himachal Pradesh, India
- Basant K. Tiwary** Department of Bioinformatics, Pondicherry University, Pondicherry, India
- Renato Hidaka Torres** Federal University of Pará, Belém, Brazil
- Sanjaya Shankar Tripathy** Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Jharkhand, India
- Aruna Tyagi** Division of Biochemistry, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
- Megha Ujwal** ICAR-National Institute for Plant Biotechnology, New Delhi, India
- Ahu Altinkut Uncuoğlu** Department of Bioengineering, Faculty of Engineering, Marmara University, Istanbul, Turkey
- Rakesh Kumar Verma** Department of Biosciences, Mody University of Science and Technology, Sikar, Rajasthan, India

Pritesh Vyas Dr. Khem Singh Gill Akal College of Agriculture, Eternal University, Baru Sahib, Himachal Pradesh, India

Ajit Kumar Singh Yadav Department of Computer Science and Engineering, North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India

Anurag Yadav College of Basic Sciences and Humanities, Sardarkrushinagar Agricultural University Dantiwada, Palanpur, Gujarat, India

Dinesh Yadav Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India

Kanchan Yadav Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India

Kusum Yadav Department of Biochemistry, University of Lucknow, Lucknow, Uttar Pradesh, India

Manoj Kumar Yadav Department of Agricultural Biotechnology, College of Agriculture, Sardar Vallabh

Bhai Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India

Pramod Kumar Yadav Department of Computational Biology & Bioinformatics, JIBB Sam Higginbottom University of Agriculture, Technology & Sciences, (Formerly AAI-DU), Prayagraj (Allahabad), Uttar Pradesh, India

Shikha Yashveer Department of Molecular Biology, Biotechnology & Bioinformatics, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India

Raphy Zarecki Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel; Department of Environmental Hydrology & Microbiology, Zuckerberg Institute for Water Research, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel

Ezgi Çabuk Şahin Department of Biology, Faculty of Science & Arts, Marmara University, Istanbul, Turkey

This page intentionally left blank

About the editors

Dr. Pradeep Sharma received his PhD from HAU Hisar. For his PhD thesis, he worked on cloning and characterization of phyto-viruses. He performed his postdoctoral research (2006–08) at Tohoku University, Japan, with Prof. M. Ikegami as JSPS fellow, where he worked on genome sequencing and RNA silencing. He also did postdoctoral research at the ARO Volcani Center, Israel, with Prof. Y. Gafni (2005–06) in nuclear import of genes using yeast one hybrid system. He worked as DST Scientist (2006) and visiting scientist at South Dakota State University (2011) for allele mining and comparative genomics in rice and Oklahoma State University (2016) on small noncoding RNAs. In 2008 Dr. Sharma joined the faculty in the Division of Crop Improvement at ICAR-Indian Institute of Wheat and Barley Research, Karnal, India, where he is a principal scientist.



He directs a research group studying the role of sRNAs and epigenetics for biotic and abiotic stresses, bioinformatics, and NGS-based marker discovery in wheat.

His group also decoded Karnal bunt genomes, development of SSR markers for population structure and diversity analysis of bunts and smut fungi, and genome-wide analysis of TFs, drought- and heat-responsive microRNAs and transcriptomics. He teaches courses in molecular biology and computational biology to graduate students. Dr. Sharma published more than 120 national and international research papers, 25 invited chapters, and 10 scientific review articles and edited 8 books on biotic and abiotic stresses, including RNAi technology. He has guided five students for PhD in biotechnology and bioinformatics as well as mentor for postdoctoral fellowships and DST-Women Scientist-A. Dr. Sharma was conferred the Young Scientist Award (biannual 2005–06) of the National Academy of Agricultural Sciences and the Pran Vohra Award (2008–09) of the Indian Science Congress Association, Fellow of National Academy of Biological Sciences (2015), Fellow of Indian Virological Society (2012), and Fellow Society for Advancement of Wheat and Barley Research (2019). He has worked at and visited many pioneering laboratories of the United States, the United Kingdom, Japan, France, China, the Netherlands, Indonesia, Turkey, and Israel.

Prof. Dinesh Yadav is a professor in the Department of Biotechnology at D.D.U Gorakhpur University, Gorakhpur, India. Presently, he is a coordinator of newly established “Centre for Genomics and Bioinformatics” at D.D.U Gorakhpur University. He was the Head of Department of Biotechnology, D.D.U. Gorakhpur University from August 2016 to August 2019. He is also serving as nodal officer of IPR cell at DDU Gorakhpur University since January 2019. He has also served as an associate professor in the Department of Molecular Biology and Genetic Engineering at G.B Pant University of Agriculture and Technology, Pantnagar, from 2006 to 2009. He has completed his Master’s Degree in biotechnology from Devi Ahilya University, Indore in 1996 and PhD from G.B. Pant University of Agriculture and Technology, Pantnagar in 2002. He has more than 19 years of teaching/research experiences. His areas of specialization are molecular biology, bioinformatics, plant biotechnology, and enzyme technology. He has published more than 130 research papers, including reviews, books, chapters in books, and proceedings in conferences and more than 230 GenBank accession numbers.



He has guided 12 students for PhD in biotechnology and presently 3 students are pursuing PhD. He has been a mentor for five postdoctoral fellowships, namely, UGC-Dr. D.S. Kothari (twice), DST-Women Scientist-A, DST-Women Scientist-B, and SERB-National PDF. He has also supervised 90 students for M.Sc. dissertation/short project work in Biotechnology. He has carried out projects from funding agencies such as DBT, UGC, UP Council of Agricultural Research, Lucknow, and DST. He has availed DST-BOYSCAST Fellowship at Australian Centre for Plant Functional Genomics (ACPFG), University of Adelaide, South Australia, from May 3, 2012 to April 12, 2013. He has been

awarded Dr. Pushpendra Kumar Gupta Vishisht Krishi Vaigyanik Puraskar-2015 in the field of agricultural sciences by Uttar Pradesh Academy of Agricultural Sciences (UPAAS), Lucknow, and Young Scientist Award-2008 by Uttarakhand State Council of Science & Technology in the discipline Biotechnology, Biochemistry and Microbiology. He is a life member of scientific societies, namely, BRSI, Trivandrum; SBC(I), Bangalore; Indian Science Congress Association, Calcutta; Society of Plant Biochemistry and Biotechnology, IARI, New Delhi; Association of Microbiologist of India (AMI), New Delhi; and UPAAS, Lucknow. He has delivered more than 80 invited talks/lectures in conference/symposium and also served as a mentor for DST-INSPIRE Science internship camp and Inspire awards. Presently, he is working on plant-specific transcription factor—DOF (DNA binding with one finger) and nuclear factor-Y (NF-Y) for developing biotic and abiotic stress-tolerant crops and pectinase group of enzymes with potential applications in different industries. His research work has been published in National and International journals of repute such as *Journal of Experimental Botany*, *Planta*, *Theoretical and Applied Genetics*, *Process Biochemistry*, *Molecular Biology Reports*, *Plant Systematics and Evolution*, *Molecular Biotechnology*, *Applied Biochemistry and Biotechnology*, *Annals of Microbiology*, *Journal of Basic Microbiology*, *3 Biotech*, *Biologia*, *Journal of Cereal Sciences*, *Physiology and Molecular Biology of Plants*, *Biochemistry (Moscow)*, *Current Proteomics*, *Interdisciplinary Sciences: Computational Life Sciences*, *Biocatalysis and Agricultural Biotechnology*, *Cell Biochemistry and Biophysics*, *Online Journal of Bioinformatics*, *Chemistry and Ecology*, *African Journal of Biotechnology*, *Applied Biochemistry and Microbiology*, *SugarTech*, *Enzyme Research*, etc.

Prof. (Dr.) Rajarshi Kumar Gaur earned PhD in 2005, now a professor in the Department of Biotechnology at Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India. His PhD was on molecular characterization of sugarcane viruses, namely, mosaic, streak mosaic and yellow luteovirus. He received MASHAV fellowship of the Israel government for his postdoctoral studies and joined The Volcani Centre, Israel and BenGurion University, Negev, Israel. In 2007 he received the Visiting Scientist Fellowship from Swedish Institute Fellowship, Sweden, to work in the Umeå University, Umeå, Sweden. He received postdoctoral fellowship from ICGEB, Italy, in 2008. He has made significant contributions on plant viruses and published 130 national/international papers, authored 18 edited books, and presented near about 50 papers in the national and international conferences.



He has honored as Fellow of Linnean Society, Fellow of Royal society of Biology, Fellow of Society of Plant Research, Fellow of Society of Applied Biology (FSAB), and Fellow of International Society of Biotechnology (FISBT). He has bagged many other awards, such as Prof. B.M. Johri Memorial Award by Society of Plant Research (SPR); Excellent Teaching Award by Astha Foundation, Meerut; UGC-Research Teacher Award; Young Scientist Award-2012 in biotechnology by Society of Plant Research (SPR), Meerut; and Scientific & Applied Research Centre Gold Medal Award-2011 for outstanding contribution in the field of biotechnology. He has visited several laboratories of the United States, Canada, New Zealand, the United Kingdom, Thailand, Sweden, and Italy. He is also associated with several international journals as an academic editors and reviewers such as *Frontiers in Microbiology*, *PLoS One*, *Scientific Reports*, *3Biotech*, *Journal of Plant Growth Regulators*, *Molecular Biology Report*, *Plant Disease*, *Current Genomic*, *Scientific African*, *Indian Phytopathology*, *VirusDisease*, etc. Currently, he is involved in many national and international grants and international collaborative projects on plant viruses and disease management.

Foreword

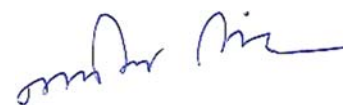
The last 10 years were considered to be a new era of bioinformatics and computational biology, which widens the pace of scientific invention and development in life science. The inclination of computer science in the area of agriculture has changed the way we usually do research related to plants in previous decades. To address the advances of bioinformatics in agriculture, the editors, Dr. Pradeep Sharma, Prof. Dinesh Yadav, and Prof. Rajarshi Kumar Gaur, have undertaken the thorny assignment of capturing the status and future trends of crop production systems. *Bioinformatics in Agriculture: Next-Generation Sequencing Era*, delivered by the proponents of Agroinformatics, offers a wealth of information about the scientific breakthroughs and discoveries aiming to meet the global challenges of the diminishing amount of arable land as well as energy shortage, malnutrition, and famine. Bioinformatics tools enable the generation, collection, and interpretation of biological data on key factors that are responsible for better crop yield. This book presents the high-throughput technology, which is used for the generating of data in the form of biological sequences that could be DNA, RNA or protein. The use of NGS has introduced a new age of omics approaches that revolutionize information generation in agriculture improvement.



The book consists of 37 chapters that are distributed in four sections: [Section I](#), Bioinformatics and Next Generation Sequencing Technologies, [Section II](#), Omics Application, [Section III](#), Data Mining and Markers Discovery, and [Section IV](#), Artificial Intelligence and Agribots.

The book highlights crop improvement such as yield enhancement, biotic and abiotic resistance, genetic modification, bioremediation, food security, etc. It explores how the different omics technology independently and collectively would be used to improve the quantitative and qualitative traits of crop plants. It explores how the different omics technologies, especially the most recent ones (proteomics, metabolomics, nutrigenomics, and metagenomics) would be used to improve the quantitative and qualitative features of crop plants. The book also discusses more efficient farming practices of recent technological advancements and solutions to current bottlenecks in farming. Application of Artificial Intelligence or machine intelligence across the farming sector is also mentioned, which could act to be an epitome of shift in how farming is practiced today. The chapters contain numerous beautiful and revealing illustrations helpful for the reader to grasp the essence of the message. Throughout the book, the approaches have been scrutinized with a critical eye as is characteristic of dedicated science professionals.

I am confident that this excellent book provides an insightful overview of the prospects and challenges of plant biotechnology, both to researchers and students in this fascinating field. It is thrilling to see editors take on this project and important topics. I hope that many readers of the book will become informed advocates of bioinformatics.



Nagendra Kumar Singh
*National Institute for Plant Biotechnology,
Pusa Campus, New Delhi, India*

This page intentionally left blank

Preface

Agricultural biotechnology is playing a significant role in developing appropriate strategies to be utilized by breeders for crop improvement programs. With an estimated world's population of 7–9 billion by 2050 and climate change, the goal of achieving global food and nutritional security will be extremely difficult by using conventional methods of agriculture. Technological innovations as the outcome of biotechnological research in the form of emerging omics-driven tools seem to have immense potential to deliver in near future. The recent revolution in genome sequencing technologies popularly referred to as next-generation sequencing (NGS) resulted in deciphering of several genomes of important crops along with model crops. With the drastic increase in the genome sequence information, its storage, retrieval, annotation, and analysis need efficient computational intervention in the form of emerging multidisciplinary science of bioinformatics. The “Science of Omics” has several subbranches but the most popular among them are genomics (structural, functional, and comparative); proteomics; and metabolomics, where efficient tools have been developed and are being applied in research.

The recent developments in agriculture need special attention among the students and researchers so that they get an insight into the relevance of technological innovations with an ultimate aim for crop improvement to sustain life. Keeping this in our mind, we thought of coming with a book which could provide all aspects of agricultural research where bioinformatics has a central role to play. We are really happy to share that we got the best contributions from experts all over the world who discussed not only the basics about the omics and bioinformatics but also the recent advances such as big data analysis, artificial intelligence, and deep learning.

The advances in biotechnology such as the NGS technologies have required the use of bioinformatics in agriculture and crop management. Computational biology manages biological data that help in decoding of plant genomics and proteomics. Bioinformatics develops algorithms and suitable data analysis tools to infer the information and make discoveries. Application of various bioinformatics tools in biological research enables storage, retrieval, analysis, annotation, and visualization of results and promotes better understanding of biological system in fullness. The exponential growth of sequencing and genotyping technology and the parallel growth of bioinformatics and online biological resources can successfully be harnessed for innovative breeding and pathogen diagnostic approaches.

In addition, we believe that this book will serve as a useful reference for both bioinformaticians and computational biologists in the omics era. The chapters will be distributed in four sections.

Section I: Bioinformatics and next-generation sequencing technologies (Chapters 1–14)

This section is devoted on bioinformatics as a central tool for the interpretation and application of biological data. Using various omics tools implemented by a wide range of programmatic languages, bioinformatics tools organize, analyze, and interpret biological information at the molecular, cellular, and genomic level which can be used for crop improvements. The combined power of NGS and bioinformatics is vital for genomics, proteomics, transcriptomics, and metabolomics that can help for the crop improvements.

Section II: Omics application (Chapters 15–26)

This section describes the application of various omics technology and their holistic approach for quantification and characterization of genes, transcripts, proteins, and metabolites. The chapter discussed the genomic studies of crop plants such as rice, maize, wheat, tomato, potato, and tea that provided the insights into total number of genes, gene organization, genetic mapping, and role of genes in various metabolic processes. Approaches of bioinformatics tools toward abiotic and biotic stresses are the part of this section.

Section III: Data mining and markers discovery (Chapters 27–33)

This part of the edited book deals with the need of utilization of information and communication technologies, which will enable the extraction of significant data from agriculture in an effort to obtain knowledge and trends. The chapters also describe the data mining and marker-based technology that provide information about crops and enable agricultural enterprises to predict trends about customer's conditions or their behavior. The need of bioinformatics of agriculture data and how data mining techniques can be used as a tool for knowledge management in agriculture should be considered by researchers.

Section IV: Artificial intelligence and agribots (Chapters 34–37)

This section overviews about the current implementation of automation in agriculture, the weeding systems through the robots and drones. The deep learning, artificial intelligence, and big data methods in agriculture are discussed along with automated techniques. The implementation of all these technologies in agriculture has brought an agriculture revolution. This technology has protected the crop yield from various factors such as the climate changes, population growth, employment issues, and the food security problems.

The book is the contribution of the renowned workers and authors who are the pioneers in the field of bioinformatics over the world. Moreover, the editors will refine the authors' views in simpler manner that can be easily understandable by the readers. This book is designed to be self-contained and comprehensive, targeting professors and scientists working on bioinformatics and its related fields, such as computational biology, genomics, applied data mining, machine learning, and artificial intelligence. This edited book will also helpful to policy makers and other stakeholders to formulate effective policy recommendations for crop improvements.

Pradeep Sharma¹, Dinesh Yadav² and Rajarshi Kumar Gaur²

¹*ICAR-Indian Institute of Wheat and Barley Research, Karnal, Haryana, India,*

²*Department of Biotechnology, Deen Dayal Upadhyaya Gorakhpur University,
Gorakhpur, Uttar Pradesh, India*

Section I

Bioinformatics and next generation sequencing technologies

This page intentionally left blank

Chapter 1

Advances in agricultural bioinformatics: an outlook of multi “omics” approaches

Nisha Singh, Megha Ujinwal and Anuradha Singh

ICAR-National Institute for Plant Biotechnology, New Delhi, India

1.1 Introduction

For an ever-increasing global population, it is vital to improve food productivity in the 21st century (Singh, Bhatt, Rana, & Shivaraj, 2020; Singh, Mahato, et al., 2020; Singh, Rai, & Singh, 2020). Plants have not only served as food but also other resources such as resin, oil, fuel, dyes, drugs, and secondary metabolites (Challam, Nandhakumar, & Kardile, 2019). Recent advances in plant biotechnology have been greatly shifted from genetically modified crops and gene manipulation to multiomics approaches. Novel approaches entail that phenomics, genomics, transcriptomics, proteomics, metabolomics, ionomics, and bioinformatics have great potential to identify and characterize the new traits in plants to meet environmental status (Lepcha, Kumar, & Sathyanarayana, 2019). Due to fast development of omics tools, not only quality, nutrition composition, and taste of food crops increase but also the agricultural production, crop protection, and agricultural economics also develop very well (Singh, Bhatt, et al., 2020; Singh, Mahato, et al., 2020; Singh, Rai, et al., 2020). The application of multiomics methods has enhanced the uniformity and predictability of plant breeding (Van Emon, 2016). Omics has also provided insight into the molecular pathways of insect pesticide resistance and plant herbicide tolerance, allowing for more effective pest management. It enables a system biology approach to work out the complicated interactions between genes, proteins, and metabolites in an interested trait/phenotype. Chemical analytical procedures, bioinformatics, and computer analysis are all used in this integrated approach to improve crop protection and improvement. It also accelerated the development of genome-scale resources in applied and emerging model plant species and boosted translational research by integrating knowledge across plant species (Mochida & Shinozaki, 2010). Generally, crop traits are typical quantitative traits, controlled by multiple genes. That's why highly throughput omics techniques are integrated with bioinformatics tools to identify the factors affecting the growth and yield of food crops (Rhee, Dickerson, & Xu, 2006). Due to next-generation sequencing (NGS) technology, crop productivity and their research field have been explored. NGS is greatly accepted in targeted genomic regions, transcriptomics, whole-genome sequencing, and low-throughput practices such as genome-by-sequencing (GBS) (Poland et al., 2012; Semba, 2016).

Interdisciplinary techniques are required for plant breeding to increase crop production and solve breeding challenges (Moose & Mumm, 2008). As a result, approaches such as high-throughput genomics, proteomics, transcriptomics, and bioinformatics are critical in increasing the production rate for enhanced crop growers in order to expedite genetic gain. This new sector has the potential to provide a platform for more precise gene functional prediction in a range of complex situations. In this chapter, we address the developments in agricultural bioinformatics and how multiomics approaches allow accurate breeding and overcome barriers to crop improvement (Fig. 1.1).

1.2 Different types of “omics” approaches

1.2.1 Phenomics

The introduction of new crop types and improved production technologies, such as contemporary irrigation methods, pesticides, synthetic nitrogen fertilizer, and other management techniques, contributed to a substantial increase in food production due to the Green Revolution in the 1960s. (Rahman et al., 2015). The recent development in phenotyping techniques

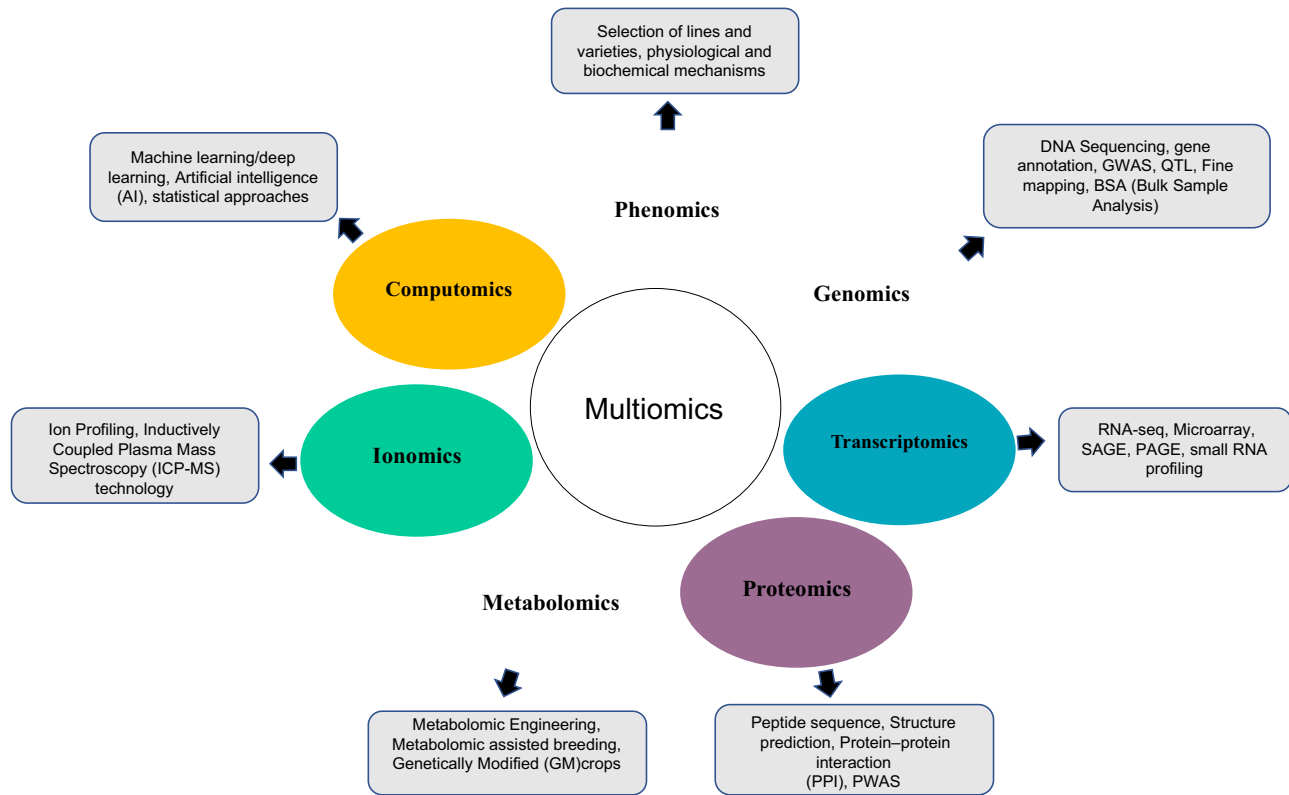


FIGURE 1.1 Various disciplines of omics research.

for plants and DNA sequencing, along with the study of massive datasets, has given rise to the term “phenomics.” Phenomics is the description of all phenotypes, ranging from molecules to organ levels, at various levels. The word “phenomics” was coined by Steven A. Garan in 1996. The term “phenomics” defines the imaging techniques that allow scientists and researchers to learn about plants at their root or whole plant level and the inner workings of the leaves. The term also refers to the entire organism research, which involves the use of the high-performance phenotyping and the data analysis in terms of development, performance, structure, architecture, and data acquisition (Pasala & Pandey, 2020). Phenomic technology can be used to research large-scale individual cells, leaves, or plants, that is, ecosystems. Phenomics is the science of the processing and analysis of large-scale phenotypic data (Heffner, Jannink, & Sorrells, 2011; Lu, Savage, Larson, Wilkerson, & Last, 2011). It is further interconnected with other “omics” technologies such as genomics, transcriptomics, and metabolomics in order to evaluate plant output in the field and link it to the core molecular genetics. High-throughput phenomics, which included imagery techniques, was used to phenotype multiple plant populations in a short amount of time (Yang et al., 2020). 3D imaging, infrared imaging, fluorescence imaging, visible light scanning, and magnetic resonance are the examples of phenomic high-throughput techniques (Sozzani, Busch, Spalding, & Benfey, 2014).

- In 3D imaging techniques, plant pots move through an imaging chamber on a conveyor system. Automatically 3D models are generated in a computer. (Tsafaris & Noutsos, 2009).
- Thermal infrared cameras use light to investigate plant-canopy temperatures in the far-infrared spectrum area from 15 to 1000 nm. The temperature rise will further help research production, salinity and drought tolerance, and photosynthesis efficiency (Nasarudin & Shafri, 2011).
- When an object refracts light at a certain wavelength while absorbing light at a different wavelength, a fluorescence picture appears. This technique facilitates the photosynthesis process and plant health measurements. Chlorophyll fluorescence is used to research the effect of various genes or environmental factors on photosynthesis performance (Baker, 2008; Maxwell & Johnson, 2000).
- In visible light scanning, a difference in color provides an estimate of the plant/leaf senescence. The senescence of matured leaves represents mechanisms of escape or avoidance adopted by the plant under conditions of water stress, whereas stay-green genotypes under water stress will continue the photosynthesis process and are known as tolerant (Howarth, Gay, Draper, & Powell, 2011).

- Magnetic resonance imaging is a type of imaging that is commonly used to analyze plant roots. The root images are taken using a magnetic field and radio waves in the same way that bodily organs are imaged in medicine (Borisjuk, Rolletschek, & Neuberger, 2012).

Study in crop phenomics incorporates agronomy, life sciences, information technology, mathematics, and engineering and integrates high-performance research (Fig. 1.2). Computing and artificial intelligence (AI) technologies in a dynamic setting are used to explore diverse phenotypic knowledge on crop growth. The ultimate objective is to develop an efficient technological infrastructure capable of high-throughput, multidimensional, big data, intelligent crop phenotyping, and automatically measuring manners (Zhao et al., 2019). After identifying the necessity for numerous traits to be phenotyped quickly and reliably, several next-generation and high-throughput plant phenotyping platforms (HTPPs) were developed to correctly measure trait values and evaluate variance between individuals (Hartmann, Czauderna, Hoffmann, Stein, & Schreiber, 2011). HTPPs have enabled better approaches to the link between characteristics, plant development, growth, and reproduction in a variety of situations (Brown et al., 2014). This leads to a better understanding of the plant’s complete phenomenon in a wide range of environmental and growth settings.

1.2.1.1 Applications

1. Abiotic stress—In different environmental conditions, drought-tolerant wheat crops are used with different quantities of water at different growth stages. Researchers have to research the productivity of crops in the field over an entire growing season to breed drought-tolerant wheat. Under drought stress conditions, phenomic remote sensing technology can measure plant growth, canopy temperature, and other characteristics. (Berger, Parent, & Tester, 2010; Chen et al., 2014; Munns, James, Sirault, Furbank, & Jones, 2010).
2. Rapid and efficient mutant screening—In the domain of phenomics, measurements can be made on multiple plants at the same time and during the course of the growing season. Phenomic approaches have been used to identify and control field disease epidemics and pathogen root assaults, as well as to screen germplasm and simulate biomass output (Miyao et al., 2007).
3. Study of various physiological processes—There are two main photosynthetic pathways of supercharging photosynthesis plants, that is, C3 and C4. Researchers in phenomics want to replace the rice C3 pathway with a more successful mechanism of C4. C4 plants may concentrate carbon dioxide within the leaf and photosynthesize more effectively than C3 plants. In Rubisco enzyme, the inefficiency of photosynthetic performance is a key limiting

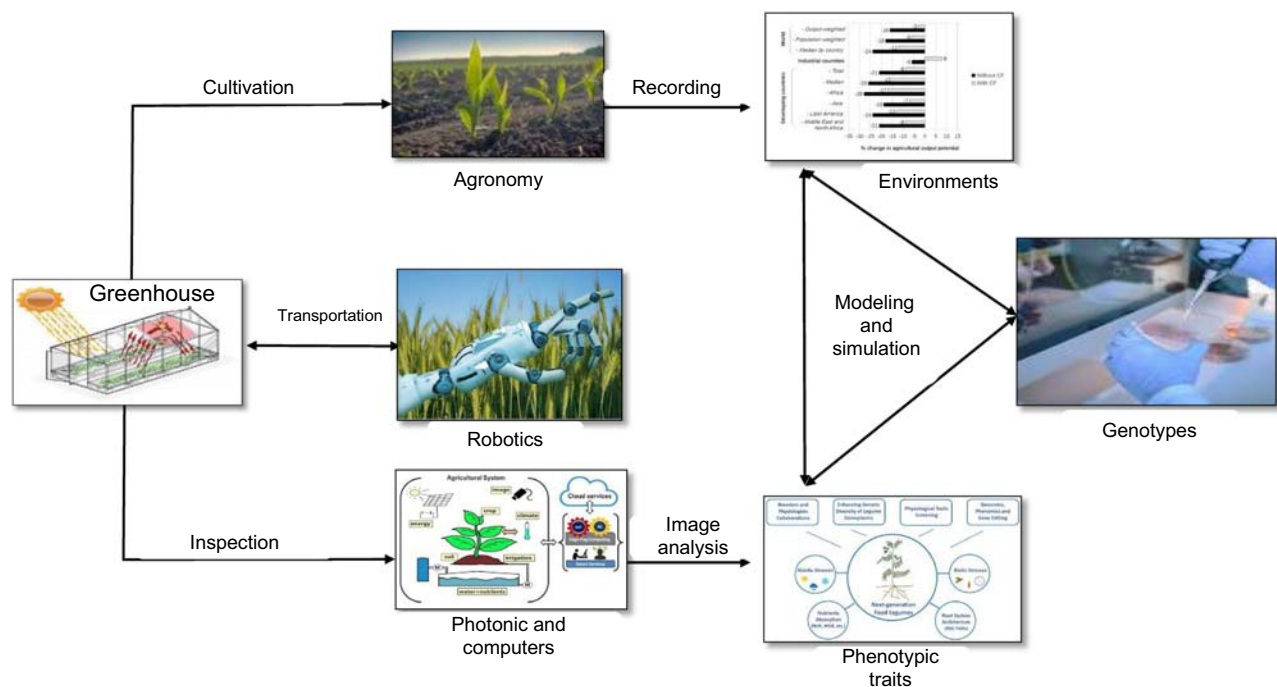


FIGURE 1.2 Steps involved in phenomic studies.

factor. Using phenomics, researchers are looking for wheat types with increased Rubisco production and photosynthetic rates that can grow well under nutrient deficit, drought, and salinity (Baker, 2008).

1.2.1.2 Challenges

Crop yield is the product of complex dynamic processes that occur between the genome, the climate, and management. In crop breeding programs, however, none of this complex knowledge is used to affect the output of a specific genotype. The challenge is to develop nondestructive methods that can be used to rapidly quantify performance traits over time and inform selection decisions on high numbers of genotypes in the field. In order to calculate crop output, agronomists and farmers often currently have to rely on challenging and damaging methods and lack the resources to track crop performance in the field. Phenomics may provide some strategies to improve the efficiency of farm-scale crop assessment (Zhao et al., 2019).

For crop morphological, structural, and physiological data, we emphasize three multicharacteristics: multidomain (phenomics, genomics, etc.); multilevel (conventional small to medium scales up to omics on a broad scale); and multi-scale (crop morphology, structure, and physiological data from cell to whole plant). The association study in the new age called “-omics” does not satisfy the single and individual phenotypic information, and the systematic and full phenomic information will be the basis for future research (Coppens, Wuyts, Inzé, & Dhondt, 2017).

1.2.2 Genomics

Hans Winkler coined the term “genome” in 1920 to describe a haploid set of chromosomes with their genes, whereas Thomas Rodrick coined the term “genomic” to describe the structure, function, and inheritance of an organism’s genome. (Griffiths et al., 2005). Genomic knowledge has provided perception into the total number of a gene, gene mapping, gene organization, and role of genes in various metabolic processes. Earlier, Sanger technology for DNA sequencing was quite expensive, time-consuming, and laborious. Innovation in DNA sequencing, that is, NGS technologies prompted a standard change in the field of genomics (Lister, Gregory, & Ecker, 2009) (Table 1.1). NGS technologies avail a widespread platform that provides deep knowledge of genomic sequences (Metzker, 2010; Pollard, Gurdasani, Mentzer, Porter, & Sandhu, 2018).

Resequencing combined with reference genome sequencing outcomes is a prominent application that fulfils the feature of NGS technologies (DePristo et al., 2011). Even polymorphisms in ecotypes and cultivars closely related to DNA polymorphisms, such as single-nucleotide polymorphisms (SNPs) and insertion–deletion polymorphisms, were classified using NGS-based resequencing (InDels).

As a result of rapid technological advances in the omics area, we need to use available genomic research for many plants of nonmodel and model species which led us to recognize another translation field of plant science, that is, plant genomics. Advances in plant genomics, huge array of *denovo* sequencing, assembly, annotations for can be easily done in nonmodel plant species. Further we can developed a costeffective genotyping technologies to enrich breeding program. For instance, *Arabidopsis thaliana*, a model plant of 125Mb, 25,489 individual genes, and 14% recurring elements, published in 2000, was the first sequenced genome for plants (UNFAO, 2015). More than 109 plant genomes, 21 monocots and 83 eudicots, 10 model and 15 nonmodel plant genomes, and 5 nonflower and 69 plant species with 6

TABLE 1.1 Different NGS technologies for genomic sequence.

Technologies	Applications	References
454 FLX (Roche)	Targeted resequencing (amplicon sequencing), metagenomics, transcriptome sequencing	Zhang, Zhang, Hu, and Yu (2011)
Hiseq (Illumina solex)	Genome resequencing, genotyping metagenomics	Bennett, Barnes, Cox, Davies, and Brown (2005)
SOLiD and PacBio RS (Pacific Biosciences)	Quantitative transcriptomics and genotyping	Eid et al. (2009)
Ion Torrent	De novo genome sequencing, target resequencing, genotyping, RNA-seq on low complexity	Rothberg et al. (2011)

NGS, Next-generation sequencing.

model crops and 15 relative wild crops were completely sequenced until 2015 (Michael & VanBuren, 2015). The processing of biopharmaceuticals and industrial compounds cannot be integrated into plants prior to the omics period. Studies of gene expression classify phenotype products of functional genes which can be used to boost the seed. The desirable phenotype can be generated faster than conventional plant reproduction by adding a particular gene to the plant or knocking down a gene with RNAi (Ahmad et al., 2012).

GWAS (genome wide association study) offers a wider view of working and interaction of genes. Progress in genome technology has allowed us to make model crops with appealing economic features. In various fields of crop biotechnology, genome sequencing, subsequent functional annotation, and molecular analysis were utilized (Yadav, Kumar, Kumar, & Yadav, 2018). SNPs are the most common type of DNA sequence variation found in human genomes. It was discovered in the genome’s coding and noncoding regions. As a result, the creation of a high-density SNP genotyping chip is critical for studying deep genetics and functional genomic applications in many crop species. These genotyping chips are extremely valuable for phylogenetic investigations, germplasm characterization, association mapping, background selection and evolutionary research, bulk segregant analysis, and the creation of high-density linkage maps. (Singh et al., 2015).

In this context, several SNP genotyping have been developed in different crops and animal species: rice (Chen et al., 2014; McCouch et al., 2010; Singh et al., 2015; Zhao et al., 2011), sunflower (Bachlava et al., 2012), soybean (Song et al., 2013), oil palm (Kwong et al., 2016), maize (Ganal et al., 2011; Unterseer et al., 2014), wheat (Wang et al., 2014; Winfield et al., 2016), and pigeonpea (Saxena et al., 2018; Singh, Bhatt, et al., 2020; Singh, Mahato, et al., 2020; Singh, Rai, et al., 2020) and chicken (Kranis et al., 2013) and cattle (Rincon, Weber, Van Eenennaam, Golden, & Medrano, 2011). Of them only two are entirely genic-SNP genotyping chips based on single-copy genes, that is, for rice “OsSN Pinks” 50K (Singh et al., 2015) and pigeonpea “CcSNPnks” 62K (Singh, Bhatt, et al., 2020; Singh, Mahato, et al., 2020; Singh, Rai, et al., 2020). It comprises multiple SNPs per gene, allowing gene-based haplotype association analysis.

In genomic applications, GWAS becomes an efficient tool for the identification of complex traits into plant genetics (Atwell et al., 2010). GWAS offers a number of advantages over traditional gene mapping methods, including the fact that it is more successful in plants than in people. In an ecological context, mapping tools can be used (i) to separate adaptive genetic variation from structured background variation, (ii) Quantitative trait loci (QTL) were first discovered in biparental crosses in plants, but they were limited in allelic diversity and chromosomal resolution. By offering better resolution, typically to the gene level, GWAS overcomes numerous drawbacks of classical gene mapping, and (iii) utilizing samples from previously well-studied groups where frequent genetic differences are linked to phenotypic variance (Brachi, Morris, & Borevitz, 2011). The objective of “agricultural genome,” through the analysis of crops or livestock genomes, is to find novel solution for the safety of the food industry, and sustainable productivity knowledge for the other aspects such as development of energy or design (Van Borm et al., 2015; Vander Vlugt et al., 2015; Wilson & Roberts, 2014).

1.2.2.1 Applications of genomic technologies

1. Genome sequencing and gene prediction—With the advancement of NGS technologies, we are allowed to predict gene functionality through comparative genomic studies. The first full genome sequencing of *A. thaliana*, discovering 25,000 functional genes, is compared with newly sequenced genomes to discover new genes by comparative genomic studies. Model and nonmodel plant species’ comparative genetics will classify an arrangement of their genes with respect to each other, which is then used to transfer knowledge from model crop systems to other food crops (Yadav et al., 2018).
2. Analysis of genetic variation and trait-specific marker mapping—As an important instrument for early detection of desired characters in the progeny, molecular markers have been identified. To access and amplify the variety of economically important traits of crop plants, knowledge of molecular markers can now be applied (Collard & Mackill, 2008). In the processing of large sequences and identification of SNP or SSR (simple sequence repeat), molecular markers are found throughout the genome, NGS technologies have made it possible (Salgotra, Gupta, & Stewart, 2014). These molecular markers have been used to produce genetic and physical maps and to classify the regions responsible for crop adaptation to different conditions of stress (Varshney et al., 2013). Based on their cosegregation, genetic maps reflect the location of markers in the linkage community. The creation of genetic maps with increased marker density has led to NGS technologies. To replace QTL mapping with association mapping, these enriched maps have been used. The QTL mapping connects a wider genomic region with specific features, but as it uses more markers, association mapping provides higher resolution. Thus, as a molecular characterization tool, association mapping is more informative and accurate.

3. Genetic improvement of crop plants—Omics studies have contributed to the advancement of agricultural science for food crop enhancement, feedstock, and environmental maintenance. Genomic sequencing and studies of gene expression have helped to classify the functional genes associated with a specific phenotype, and this information may be used by incorporating genes or posttranscriptional gene silencing to boost crop plants (Ahmad et al., 2012). The development of functional foods such as drought-tolerant maize, higher grain-producing rice (Ashikari et al., 2005), and bananas with longer shelf life has been made possible by genomic technologies (Mehrotra & Goyal, 2013). Plants are subjected to mutagenic reagents, popularly known as mutation breeding, for the development of designer crops with desired economic traits (Fig. 1.3). Marker-assisted breeding has chosen the progeny with the ideal character. To boost agricultural crops, molecular markers such as SSR and SNPs discovered by genome sequencing techniques have been applied (Salgotra et al., 2014).

1.2.2.2 Challenges of genomics in agricultural field

Agriculture has substantial problems in exploiting the deluge of genomic data from various sources and formats for crop development, such as the assembly of long reads of genomic sequencing and the presence of highly repetitive DNA in the plant genome sequence (Hu, Scheben, & Edwards, 2018). The gaps in the genome sequence will cause inaccuracies in the final draught sequencing. Polyploidy and heterozygosity in agricultural crops provide difficulty during the construction of their sequences. The functional annotation of numerous genes discovered has yet to be completed (Yadav et al., 2018)

1.2.3 Transcriptomics

The “transcriptome” is defined as “a complete complement of mRNA molecules formed by a cell or cell population.” The term was coined by Charles Auffray in 1996 (McGettigan, 2013). The analysis of RNA profiles within the cells at a given point in time is “transcriptomics.” In addition to RNA coding, cells often have large non-RNA coding sequences. Because of its importance, it is not as straightforward as studying the transcriptome of a cell or its complexity. However, the recent advancement of transcriptomic technology has allowed the transcriptome of a living cell to be characterized and untie the molecular base to strategically increase the development of crop plants (Pandit, Shah, & Husaini, 2018). DNA transcribing genetic information into RNA and RNA translated to protein. The core dogma of molecular biology is focused on various aspects of biological functions of cells, tissues, and species, where RNA itself is the main player for mediating the expression of genes and proteins. Thus RNA plays an important role in transcribing the DNA message (Pertea, 2012).

Transcriptomics, also known as expression profiling, is a study of mRNA expression levels in a specific cell population and provides information on expressed sequence tags (EST) in a specific tissue at a certain time. Since it is primarily a depiction of the genes which actively expressed under different conditions at any given time, and the same gene can generate many transcripts due to alternate splicing, transcriptomic is a dynamic, except in the case of mutation,

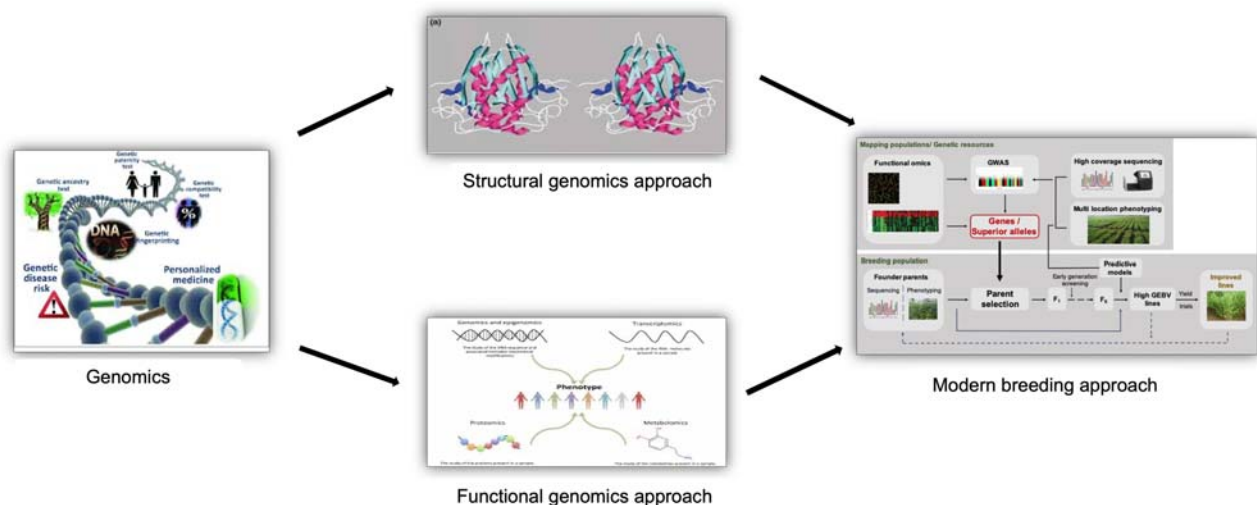


FIGURE 1.3 Different regulation of genomics used in agriculture.

unlike genome, which is approximately fixed for a specific cell line (Van Emon, 2016). Transcriptomics explore the way gene expression patterns change due to inner and external influences such as biotic and abiotic stresses (Valdés, Ibáñez, Simó, & García-Cañas, 2013). Advancements in NGS technologies have made it possible to obtain cost-effective, useful transcriptome assemblies for gene annotation (Mochida & Shinozaki, 2010). Analysis of transcriptome assemblies provide information on different functional markers related to stress-resistant response such as SSR and SNPs (Aharoni & Vorst, 2002). After acquiring the qualitative counts of each transcript, differential gene expression might be examined by normalizing the data with the use of statistical modeling. (Lowe, Shirley, Bleackley, Dolan, & Shafee, 2017). The transcriptome can now be defined using NGS technology due to RNA sequencing (RNA-seq), and the number of research utilizing RNA-seq has continuously expanded, eventually covering the microarray-induced bias (Yu & Lin, 2016).

1.2.3.1 Applications

1. Transcriptome analysis provides an important forum for examining the relationship between genotype and phenotype, providing a better understanding of underlying pathways and mechanisms that regulate cell fate and development and progression of diseases (Ruan, Le Ber, Ng, & Liu, 2004).
2. In order to understand the variation in transcriptome data during seed germination, growth, development, and different stresses, the microarray technology was used favorably (Poole, Barker, Wilson, Coghill, & Edwards, 2007).
3. As gene silencing methods for the refining of agricultural crops, practical techniques such as RNA interference (RNAi), mutagenesis, and epigenetics can be applied.
4. QTL has been mapped on crop genome related to grain development, resistance to biotic and abiotic stresses, and have been successfully applied for crop variety improvement (Saha, Sarker, Chen, Vandemark, & Muehlbauer, 2010).
5. The significance of transcriptome analyses has made it possible for relevant research groups to handle and make these data available to researchers to help them to unlock and analyze particular transcription activity at specific developmental stages of different genes. The characterization and quantification of the transcriptome was accelerated by NGS, which also strengthened the developmental evolution of advanced bioinformatics tools (Afzal et al., 2020).
6. Transcriptomics from multiple species can help researchers better comprehend complicated plant–microbe interactions. Transcriptomics can be used to improve marker discovery, the relevance of resources generated for related species, and the characterization of genes involved in various plant processes (Schenk, Carvalhais, & Kazan, 2012).

1.2.3.2 Different transcriptomic techniques with their application

1. NGS-based RNA sequencing (RNA-seq) is a method that can use NGS to analyze the sum and sequence of RNA in a sample. RNA-seq lets us investigate and discover the transcriptome, and then we can connect the genome information to the functional expression of the protein (Ozsolak & Milos, 2011).
2. We can record transcriptional profiles in each cell type using single-cell transcriptomic methods to uncover the genetic foundation of their identity and function. This knowledge of cell type-defining gene networks is important for both fundamental science and the production of crops that are more resilient to climatic and other environmental challenges. (Rich-Griffin et al., 2020).
3. DNA microarray used to study circadian clock, plant defense, environmental stress response, and fruit ripening (Aharoni & Vorst, 2002).
4. EST are used for premicroarray design.
5. SAGE (serial analysis of gene expression) used for expression analysis plants with less characterized genomes (Velculescu, Vogelstein, & Kinzler, 2000).
6. Long SAGE, a derived transcriptome used for annotation of expressed gene (Saha et al., 2010).
7. MPSS (Massive Parallel Signature Signaling) used to identify and quantify RNA transcript (Brenner et al., 2000).

1.2.3.3 Challenges

Other technologies, such as microarray hybridization, are typically regarded as inferior to RNA-seq. Due to the small amount of raw genetic material, single-cell data is constrained by low sequencing coverage and strong amplification bias. Furthermore, due to the vast genome scale, extremely repetitive areas in plant genomes, entire genome duplications, and large numbers of gene families, it is difficult to evaluate computational results (Yuan, Bayer, Batley, & Edwards, 2017). The alignment of reads to a reference genome was the first major problem posed by the advent of RNA-seq (McGettigan, 2013). While in RNA-seq, there are only a few steps that involve several stages of manipulation during the development of cDNA libraries, which may complicate its use in all forms of transcript profiling. The study

of RNA-seq outcomes is also complicated by certain manipulations during library construction. RNA-seq faces many computational challenges, including the creation of successful methods for storing, retrieving, and processing large quantities of data, which must be resolved in order to minimize image analysis and base-calling errors and eliminate low-quality reads (Wang, Gerstein, & Snyder, 2009). To analyze the huge amount the data, we don't have high-throughput machine learning (ML) algorithm to cope with this. Many researchers have found large amounts of data from RNA-seq technologies for transcriptome profiling, but we still don't have to analyze it properly by comparing it with other information (Rich-Griffin et al., 2020). On the other hand, MiRNAs induce gene silencing in plants by cleaving target mRNA or repressing translation. Although most miRNAs' biological roles are unknown, research has revealed their involvement in several developmental stages, signal transmission, disease resistance, nutritional value, and metabolomic technologies in genetic engineering (Challam et al., 2019).

1.2.4 Proteomics

The “proteome” can be identified as a cell's overall protein content that is characterized at a specific time in terms of its position, interaction, posttranslational modification, and turnover. In 1996 Marc Wilkins first used the word “proteomic” to denote the “protein complement of a genome.” The proteome characterizes much of the functional details of genes (Aslam, Basit, Nisar, Khurshid, & Rasool, 2017). To maintain structure and important regulatory function, the genome code for the protein is needed (Souda, Ryan, Cramer & Whitelege, 2011). Proteomics is the study of amino acid sequences and posttranslational modifications in order to determine their relative concentrations (Barbier-Brygoo & Joyard, 2004). In contrast to genomics, it is complex in nature subject to translational and posttranslational modification (Natarajan, Xu, Bae, & Bailey, 2007). Proteomics is a cutting-edge approach for deciphering a tissue's protein profiling in order to identify molecular entities that may be modified to generate superior crop breeds that are resistant to both biotic and abiotic stresses. (Singh et al., 2015). It has emerged as an essential tool for crop improvement as it describes the position of protein within cells that maintain homeostasis, are involved in cell signaling pathways, and are necessary for structural maintenance.

Several attempts have been made to analyze the differential proteome map of crop plants in response to a variety of stresses, including hazardous abiotic and biotic factors such as metal salinity, flooding ultraviolet-B radiation, and disease infection (Aghaei, Ehsanpour, & Komatsu, 2008; Zhen et al., 2007). The most insensitive proteomic research was done on the model plant species *A. thaliana* and rice, especially after *Arabidopsis* and rice genome decoding was reported in 2000 and 2002, respectively (Goff et al., 2002; Kaul et al., 2000). This is because protein recognition is only possible using genomic knowledge, this approach is known as proteogenomics. Similarly, the growing number of crops studied using a proteomic approach, such as rice, maize, wheat, barley, chickpea, pigeonpea, soybean, and date palm, has increased with increasing genomic DNA and EST sequencing data deposited in the public domain.

Different protein atlas was developed in different plant species. Protein atlas or expression atlas offer information on gene and protein expression in plant samples of various cell types, organism sections, developmental stages, diseases, and other factors. Atlas comprises 389 experiments investigating plants in 11 species (<http://www.ebi.ac.uk/gxa/plant/experiments>), including 7 baseline studies disclosing expression in tissues, strains, and cultivars, for example, rice, wheat, maize, and *Arabidopsis* (Petryszak et al., 2016). Many large-scale research works have now been conducted to investigate the molecular mechanisms of symbiosis between legume models and *Medicago truncatula*. Furthermore, the recently discovered genome sequence of *M. truncatula* significantly expanded the gene pool (Young et al., 2011). *Sinorhizobium meliloti* is associated with *M. truncatula* quantitative atlas of protein expression (<https://mtgea.noble.org/>). This proteome atlas contains information on 23,013 protein groups, 20,010 phosphorylation sites, and 734 active lysine acetylation sites. Using this resource, a subset of proteins with organ-specific regulation was identified. A symbiosis-specific regulation network was generated by using this putative protein atlas (Marx et al., 2016). The *Glycine max* Seq-Atlas incorporates RNA-seq data from a range of tissue collections and offers new methods for analyzing large sets of transcriptome data collected from NGS. This was possible by uniquely mapping short read sequences in RNA-seq digital gene expression analysis of paleopolyploid soybean genome. The Seq-Atlas of *G. max* (<http://www.soybase.org/soyseq>) incorporates RNA-seq data from a range of tissue collections and offers new methods for analyzing large sets of transcriptome data collected from NGS (Severin et al., 2010).

1.2.4.1 Applications

1. In order to unravel the expression of allergens in transgenic plants and to compare allergens between cultured and wild forms, the proteomic techniques (Fig. 1.4) has been used (Natarajan et al., 2007) and also it has been used for

the investigation of gene silencing materials in transgenic plants. Substantial suppression of GlymBd 30K, a dominant soybean seed allergen, was confirmed by reverse genetic method (Herman, Helm, Jung, & Kinney, 2003).

- Quantitative proteome investigations using high-resolution and mass-precision tools have added to our knowledge of plant growth, development, and interactions with the environment. This capability is especially beneficial for crops because it can help with not just increasing nutritional value and yield but also understanding crop adaptation mechanisms in response to abiotic challenges (Hu, Rampitsch, & Bykova, 2015).
- Translational plant proteomics is a proteomic extension from expression to functional, structural, and finally, the translation of ideal characteristics and their manifestation. The findings of proteomics for foods by translational proteomics are possible to apply authenticity, food security and protection, sustainability of resources, human health, improved economic standards, and environmental management (Agrawal et al., 2012).
- To increase the photosynthetic efficiencies of crop plants and their resistance to abiotic stress, C4 plants have been found to produce two forms of chloroplasts and are thus more efficient in terms of energy conversion. A comparative proteomic analysis was conducted with C3 chloroplast plants and C4 to classify the proteins that are responsible for more successful light fixation (Zhao, Chen, & Dai, 2013) (Fig. 1.4).

1.2.4.2 Technologies involved in proteomic analysis

- The most frequent gel-based approach used in a proteomic laboratory for separating the protein portion of the cellular extract is two-dimensional electrophoresis, which is reasonably easy and inexpensive (Xu, Xu & Huang, 2008) (Fig. 1.5).
- Electrospray ionization is used to convert peptides into ions by passing them through high-voltage columns. In mass spectrometry, time of flight (TOF) is a methodology for analyzing the mass of peptide ions. The most extensively used Ms (mass spectroscopy) technique is matrix-assisted laser desorption/ionization TOF. (Kersten et al., 2002).
- Ms-based proteomics can be utilized for protein profiling, recognition, and quantification, as well as the investigation of protein changes and interactions. (Aebersold & Mann, 2003) (Fig. 1.4).
- iTRAQ (isobaric tags for relative and absolute quantification) proteomic study has been conducted in the quantification of protein, best suited for impartial untargeted biomarker discovery and the quantification of protein acetylation in HCT (Helminthosporium carbonum toxin)-treated or pathogen-infected plants. These studies reveal that HCT plays an important role in altering activity of histone deacetylases, which further influences both histone and

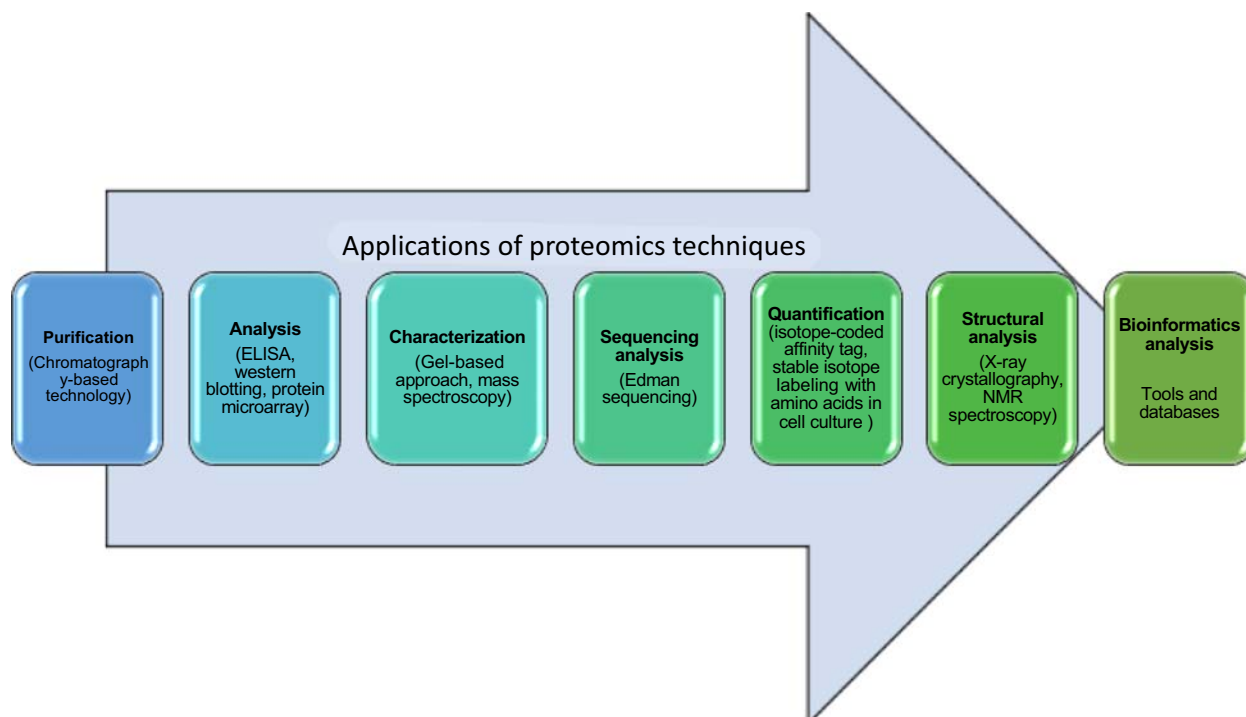


FIGURE 1.4 Application of proteomic techniques.

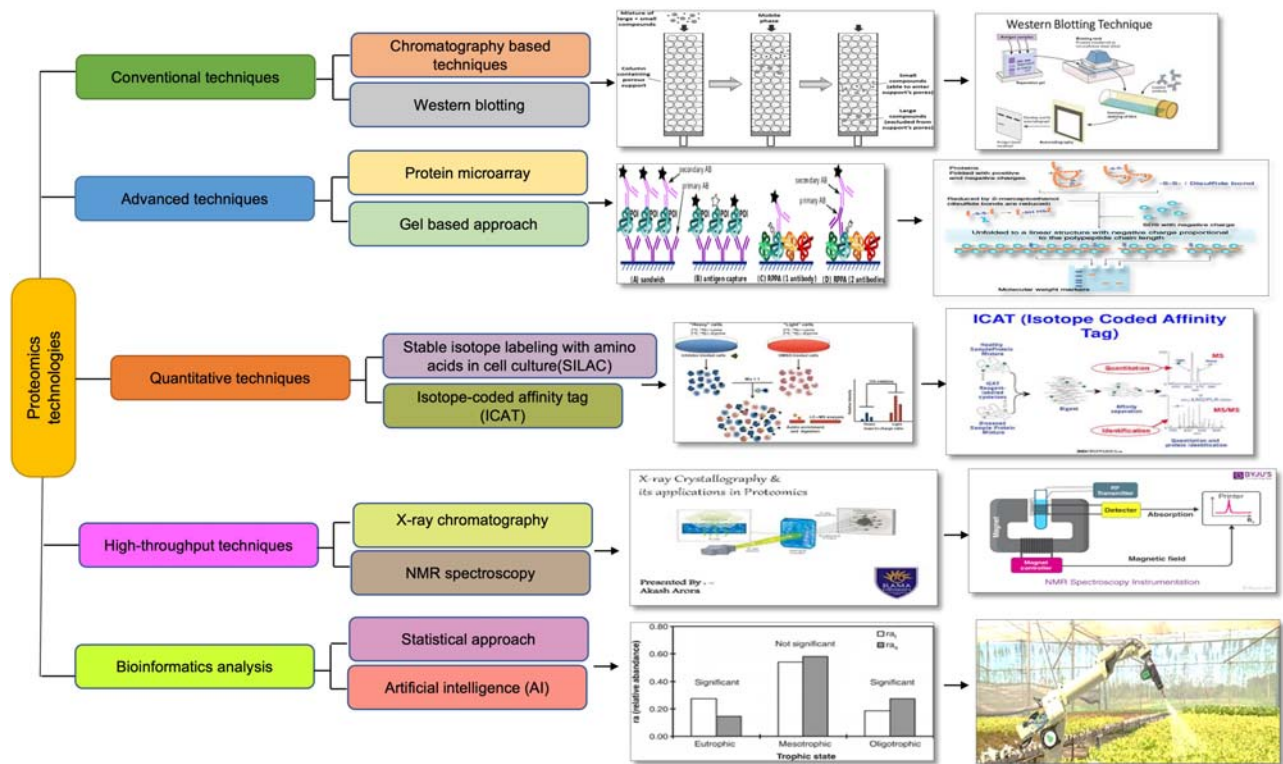


FIGURE 1.5 Overview of proteomic techniques.

nonhistone protein during plant pathogen interaction. This approach is used for functional annotation and enrichment analysis, clustering analysis, network analysis, and statistical analysis (Walley, Shen, McReynolds, Schmelz, & Briggs, 2018).

1.2.4.3 Challenges of proteomic approaches

The samples extract abundant amounts of proteins, which would hinder the analysis of the desired protein. Proteomic analysis and data interpretation techniques do not currently have appropriate guidance available. Biological protein differences are responsible for the lower reproducibility of results from proteomics. Therefore, under regulated conditions, the research should be carefully performed. Proteomics also relies on protein-function prediction instruments and software in silico. Protein functions are also predicted by homology quests with open datasets, which may lead to incorrect predictions (Yadav et al., 2018; Gong & Wang, 2013).

1.2.5 Metabolomics

Metabolomics is a new method based on finding out the essence of dynamics and biochemical structure within the living system (Dixon et al., 2006). The metabolite is the end of cellular regulatory processes, and its level is also seen as a definite response of the biological system to changes in genetics and the environment. In the form of environment–gene interaction, mutant characterization, marker identification, and drug discovery, metabolomics stands out significantly (Razzaq, Sadia, Raza, Khalid Hameed, & Saleem, 2019). Metabolomic strategies have the ability to optimize agricultural product trait production and biorefining, that is, the plant-based economy (Dixon et al., 2006).

Plants generate more than 20,000 metabolites that are involved in many resistance and stress tolerance responses and play a key role in enabling the adaptation of unique ecological niches and contributing to the color, taste, aroma, and fragrance of fruits and flowers (Oksman-Caldentey & Saito, 2005; Bino et al., 2004). The customs of agricultural varieties range from obsolete foodstuffs to foodstuffs with certain useful features, such as nutritional values and consumer products derived from fibers, latex, packaging materials, polymers, and certain essential chemical fuels (Abbas & Cheryan, 2002). The metabolomic approach in agriculture seeks to understand the biology of metabolites and apply that knowledge to food and environmental safety.

Many metabolomic extraction and analysis approaches are employed to determine the complicated nature of the metabolite and its diverse chemical composition (Wishart, 2011). Integration with metabolomics of modern plant genomic instruments, databases, and bioinformatics tools (GBS, genome-wide genetic variants and whole-genome sequencing) (Table 1.2) reveals an exciting horizon for crop improvement (Zivy et al., 2015). The metabolomic technique performs metabolic profiling of biofluid and various cell tissues to represent the whole physiological makeup of the cell (Yang et al., 2018). The metabolome is made up of several various chemical and physical components, such as pka, stability, molecular weight, size, polarity, and solubility. (Villas-Boas, Koulman, & Lane, 2007). Various analytical technologies have been used for these chemicals to be isolated, detected, and quantified. The metabolite content in agriculture is linked to a variety of processes, including fruit development, resistance to adverse environmental circumstances, stress tolerance, and pathogen infection. These substances are analyzed using a variety of analytical methods. For example, a wide variety of compounds, such as vitamins, coenzymes, carbohydrates, amino acids, and many more, can be analyzed by liquid chromatography (LC) combined with mass spectrometry (Carreno-Quintero, Bouwmeester, & Keurentjes, 2013).

1.2.5.1 Metabolomic application in crop production

The content of metabolites is linked to processes of development and differentiation, processes of fruit maturation, resistance to adverse environmental factors, stress-related issues, and pathogen attacks, especially in agriculture, among others. Some applications are:

1. As plants are capable of generating different chemical compounds, successful engineering of plant metabolic pathways associated with modern biotechnology would be beneficial to humankind (food and medicines) (Oksman-Caldentey & Saito, 2005). Knowledge-based approaches to metabolic engineering will help to continuously increase the input and output of engineering plants by inculcating large datasets and logical metabolic pathway models through large-scale processing and mining of multiple omics data (Farre, Twyman, Christou, Capell, & Zhu, 2015). Vintages of endogenous sugars, for example, such as higher level sugars and simple sugar derivatives, have been

TABLE 1.2 Bioinformatics databases and tools for multiomics approaches.

Database	Purpose	URL	References
Phytozome	Comparative genomics	https://phytozome.jgi.doe.gov/pz/portal.html	Goodstein et al. (2012)
GRASSIUS	Coregulation and comprehensive collection of transcription factors	http://grassius.org/grasscoregdb.php	Yilmaz et al. (2009)
RiceNetDB	Genome-scale multiple level network reconstruction and comprehensive rice genome annotated information	http://bis.zju.edu.cn/ricenetdb/	Liu et al. (2013)
Expression atlas	Gene expression and biological conditions	https://www.ebi.ac.uk/gxa/plant/experiments	Petryszak et al. (2016)
Gramene	Comparative functional genomics	http://www.gramene.org/	Tello-Ruiz et al. (2016)
XCMS	Raw data can be entered directly an online bioinformatics application, which may then be utilized for statistical analysis and data processing	https://xcmsonline.scripps.edu	Montenegro-Burke et al. (2017)
METLIN	Used for stress response metabolic profiling in plants	https://metlin.scripps.edu/	Smith et al. (2005)
MetaGeneAlyse	Implementation of regular clustering technique, that is, ICA (independent component analysis) and k-mean	http://metagenealyse.mpimp-golm.mpg.de/	Daub, Kloska, and Selbig (2003)
MeltDB	A web-based platform for data assessment, processing, and statistical analysis that has been used in plant metabolomics	https://meltdb.cebitec.uni-bielefeld.de	Kessler et al. (2013)

successfully enhanced by discovering sugar biosynthesis and accumulation pathways by plant metabolic engineering (Patrick, Botha, & Birch, 2013).

2. Biopesticides have many benefits in agriculture, but their use is very limited due to unreliable manners, efficiency, shelf life, and restrictions on the climate (Babalola, 2010). To increase this, we need a new method, such as metabolomics, which describes the need for stimuli or gene expression to synthesize metabolites that have already been discovered. Therefore metabolomics will help to discover new metabolites and consistent biopesticides for agricultural purposes with the molecular method of gene sequencing and detection (Mishra & Arora, 2018).
3. It deals with the study of plant biochemical relationships of plants through the distinct structure of time (habitat life time to time of generation) and space (distance between habitat patches). This technique allows us to evaluate the interaction of abiotic factors with intra-interspecific interactions and multiple impacts between two trophic stages. The influence of abiotic and biotic stresses on any biochemical process through metabolite recognition is encountered in response to environmental factors. feedback (Garcia-Cela et al., 2018).
4. For phenotypic and genomic assortment, crop breeding relies on genetic markers. This, however, presents a significant problem due to marker effects for picking complicated features that frequently differ between populations. This can be overcome by using a mix of metabolomics and other omics to provide detailed information on crop plants in a larger scale context. These mQTL and mGWAS data help us to analyze the nature of interest characteristics in quantitative terms (Langridge & Fleury, 2011). Plant metabolic technologies may thus contribute to the development of more logical models linked to accurate metabolites or pathways associated with yield or quality characteristics by providing information on the number of identified metabolites that are also correlated with agronomically significant characteristics. In particular, current efforts to better understand the metabolic response to various stresses suggest that metabolomic assisted breeding could support in the development of more stress-resistant crops (Fernie & Schauer, 2009).
5. The design of the biochemical network was carried out by evaluating the relative metabolite profiles. A comprehension of the regulatory network and association of genetic material with phenotypic characters was implied by the integration of metabolome and transcriptome data (Urano et al., 2009).

1.2.5.2 Challenges of metabolomic technologies

Metabolite applications as biomarkers are constrained by the difficulties of traceability to particular pathways. Unknown metabolites have often been found during the study of LC-MS, which cannot be used for any analysis. The data produced by the study of metabolomics is vast and complicated, requiring multivariate analysis techniques. Biological factors can lead to the problem of evaluating a metabolite associated with a specific phenotype. Much of the metabolite is part of many pathways, so analyzing the metabolite linked to unique pathways of biosynthesis becomes challenging (Yadav et al., 2018).

1.2.6 Ionomics

Micronutrient deficiency (e.g., iron, zinc, and calcium) is commonly found in both developing and developed countries accounting for nearly 2 billion people (Tulchinsky, 2010). The majority of those changes rely on staple crops, including wheat, rice, and maize for survival. Mineral enrichment, or biofortification (genetic augmentation) of staple food crops, has thus been proposed as a long-term solution to the problem of mineral shortage. (Singh, Bhatt, et al., 2020; Singh, Mahato, et al., 2020; Singh, Rai, et al., 2020). Mineral concentration in these tissues is influenced by a variety of factors, including soil mobilization, root absorption, plant transport and redistribution, seed import and accumulation, and so on. Plant ionomics could be a good way to look into the link between gene(s) and ion transport and accumulation in this case. However, in comparison to other “omics” approaches, ionomics is usually in the onset because the bulk of genes and gene networks involved in ionome regulation are yet unknown. The term “ionome” refers to the examination of all mineral nutrients and trace elements found in a living organism (Salt, Baxter, & Lahner, 2008). The complex network of components, managed by plant physiology and biochemistry, is ultimately regulated by genetic and environmental factors (Baxter, 2009). Plant ionomics is the foundation for combining metabolomics and mineral nutrition. It all started with Robinson and Pauling’s belief in the late 1960s and early 1970s that an organism’s metabolite profile indicates its physiological status and provides a rich source of information (Marschner, 2011). Since several reliable technologies have been developed to simultaneously examine living beings’ metabolites and mineral nutrient components, bioinformatics and other genetic instruments, such as sequencing, genomes, and DNA microarrays, may be used to compare Robinson and Pauling’s early ideas on metabolomics with mineral ions (Lahner et al., 2003).

Mineral acquisition, distribution, and storage in plants is a complicated process requiring numerous molecular components such as transporters, channels, chelators, and some specific genes that encode and manage them (Gilroy & Jones, 2000). For plant ionomics, measurement of the composition of ions and elements of the entire plant, tissue, and even a single cell is needed. These can vary with the elements to be calculated, sample size availability, sample throughput, range of dynamic quantification, sensitivity, reliability, and accuracy.

All strategies are based on knowledge available in literature, clustered into two categories:

1. Techniques based on elements’ electronic properties:
 - a. Atomic absorption spectrometry (AAS)—In AAS, free atoms are in a gaseous state and absorb light in the form of optical radiation in order to detect chemical elements in a sample quantitatively (L’vov, 2005).
 - b. Ion beam analysis (IBA)—The IBA is a collection of modern and efficient methods for the quantitative determination of the sample elements. In IBA, a beam of accelerated charged particles traveling from the target material at a very high speed strikes the sample material, which further results in the release of particles or secondary radiation from the target material as c-rays and X-rays (Smit, 2005).
 - c. X-ray fluorescence (XRF) spectroscopy—XRF is also a reliable method for determining chemical components and concentrations in liquid or powdered (solid) materials and it has the added advantage of being a nondestructive analytical tool (Akbaba, Sahin, & Turkez, 2012).
2. Techniques based on elements’ nuclear properties:
 - a. Neutron activation analysis—It is a useful technique for determining the elemental composition of diverse materials in local environmental research (Galinha et al., 2011)

1.2.6.1 Applications of plant ionomics

1. Ionomics is utilized to investigate the process of mineral transport in plants by identifying potential transporter genes and additional functional validation. It entails using high-throughput elemental analysis technologies and merging them with bioinformatics and genetic tools (Baxter, 2009)
2. People are also using ionic data for phylogenetic analysis of plant species (White & Broadley, 2009).

1.2.7 Computomics

Computomics was developed in 2012 by Detlef Weigel, a German-American scientist and MEGAN (MEtaGenomics Analysis tool to advance the knowledge of metagenomics datasets) author, so that benefit of ML algorithms can be profited by others. In many national publications, Computomics has been featured since it is one of the very few companies focused on plant breeding and study of plant genomes. The diversity of biological life is unlocked by applying AI to genetics, phenotypes, microbiomes, and environmental datasets. Computomics is a team of world-leading ML, plant science, and bioinformatics specialists. Our advanced ML techniques enable plant breeding, agricultural, biotech, and microbiome researchers to quickly understand genomic data. Agri-technology and precision farming, today commonly referred to as digital agriculture, are new scientific fields that use data-intensive methodologies to drive agricultural productivity while reducing its environmental impact. The data generated in modern agricultural operations comes from a variety of sensors, allowing for a better understanding of the operating environment (the interaction between complex crop, soil, and weather conditions) as well as the process itself (machinery data), resulting in more precise and faster decision-making (Kong et al., 2007; Mackowiak et al., 2015).

ML and deep learning have arisen in association with big data technologies and high-performance computing to create new opportunities for unraveling, measuring, and understanding data-intensive processes in agricultural operating environments (Wang, Cimen, Singh, & Buckler, 2020). For association studies and crop improvement, measuring the functional and structural aspects of a plant phenotypic is also significant. As genomic research and sequencing technologies improve, an increasing demand for plant phenotypes to understand genomic data is emerging (Liu et al., 2014). Robotic elevated phenotyping may now be produced thanks to advances in measurement technology (high-throughput images and automated sensors) and ML. This overcomes the constraints of traditional human-based phenotyping by permitting quick production of phenotypic features and characteristics across vast populations (Singh, Ganapathysubramanian, Singh, & Sarkar, 2016). Phenotyping using ML has been used in stress phenotyping and disease control. A real-time ML-based high-throughput phenotyping methodology was developed to determine the extent of iron deficiency chlorosis in a total of 4366 soybeans from representative canopies (Naik et al., 2017). Polyploid genome assemblies with significant redundancy can benefit from ML. Highly redundant genomes are difficult to assemble using a non-ML-based assembly method that uses a linear approach to assemble repetitive sequence regions (Brenchley et al., 2012). To overcome this limitation, an

ML approach was utilized to detect assembly errors and construct high-quality bread wheat (*Triticum aestivum*) assembly. The RNA-seq mapping method also uses ML to delineate between natural and artificial splicing junctions, which has benefited in the annotation of the bread wheat genome (Mapleson, Venturini, Kaithakottil, & Swarbreck, 2017).

The most prevalent class of variations in plant genomes are SNPs (Rafalski, 2002). However, the discovery of SNP in polyploid plants remains a problem (Flint-Garcia, Thornsberry, & Buckler, 2003). SNP-ML, a ML -based analysis tool, employs neural networks and tree bagging models to effectively filter false positive SNPs. They demonstrated that SNP-ML could accurately detect SNP variants and identify real SNPs in simulated SNP variant data of peanut, cotton, and strawberry (Buggs et al., 2012; Clevenger, Korani, Ozias-Akins, & Jackson, 2018). Accelerator ML has proven to be useful in the agricultural sector and is expected to play a growing role in the improvement of plants (Van Emon, 2016).

1.2.7.1 Applications

1. The type of soil and the nutrition of the soil play an important role in the type and quality of the crop being cultivated. The quality of the soil is deteriorating because of rising deforestation, and it is difficult to assess the quality of the soil. An AI-based application called Plantix has been developed by a German-based technology that can detect nutrient deficiencies in soil, including plant pests and diseases, by which farmers can also get an idea of using fertilizer that helps improve the quality of harvest (Coopersmith, Minsker, Wenzel, & Gilmore, 2014).
2. AI-enabled technologies predict forecast weather conditions, analyze crop sustainability, and assess farms for the presence of diseases or pests and poor plant nutrition on farms with data such as temperature, precipitation, wind speed, and solar radiation, by using ML algorithms in combination with images collected by satellites and drones (Morellos et al., 2016).
3. ML methods, such as linear regression, support vector machine regression, decision tree regression, and K-nearest neighbors, have been utilized to produce hydrogen utilizing biomass gases. To evaluate the rainfall parameters in support of agriculture, decision tree, Bayesian, neural network, and random forest are applied (Jude Immaculate, Evanzalin Ebenanjar, Sivaranjani, & Sebastian Terence, 2020).

1.2.7.2 Challenges

Agriculture has been addressing major problems such as lack of irrigation system, climate rise, groundwater density, food shortage and waste, and much more. To a great degree, the fate of cultivation depends on the acceptance of different cognitive solutions. Applications need to be more robust in order to explore the vast scope of AI in agriculture. Only then it will be able to navigate regular changes in external circumstances, promote decision-making in real time, and make use of the required framework/platform to effectively collect contextual data (Slaughter, Giles, & Downey, 2008). Farmers, on the other hand, are adapting to changing circumstances by incorporating AI into their farming operations. It's just one example of how AI is revolutionizing agriculture, a growing trend that will help usher in a new era in agriculture. We'll have to be more resourceful this time around (Talaviya, Shah, Patel, Yagnik, & Shah, 2020).

1.3 Conclusions and future prospective

The advent of multiomics technologies has greatly increased our ability to feed a hungry world, especially nonagricultural regions. The various approaches discussed earlier provide useful tools that, when used together, enable for addressing the underlying process while passing through several levels of information. Through the advances made in the arena of omics, a high-throughput phenotyping platform to measure various phenotypic traits such as image-based computer vision phenotyping, image processing, and data extraction tools will be highly efficient. Integrating phenomic data with other multiomics data from genomic, transcriptomic, proteomic, metabolomic, and other physiological studies is enabling a systems biology approach for understanding plants from the single cell to the mature plant, not only during development but also under changing environmental conditions. It gives detailed information on the regulatory mechanism in response to an external stimulus at many subcellular organization levels. Despite the fact that there is a growing number of plant research using specific omics approaches to identify important biomolecules. We can see that in near future omics can revolutionize agricultural research in many exciting areas and meet the projected food demand of rising global population.

References

Abbas, C. A., & Cheryan, M. (2002). Emerging biorefinery opportunities. *Applied Biochemistry and Biotechnology*, 98, 1147-1147.

- Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928), 198–207.
- Afzal, M., Alghamdi, S. S., Migdadi, H. H., Khan, M. A., Mirza, S. B., & El-Harty, E. (2020). Legume genomics and transcriptomics: From classic breeding to modern technologies. *Saudi Journal of Biological Sciences*, 27(1), 543–555.
- Aghaei, K., Ehsanpour, A. A., & Komatsu, S. (2008). Proteome analysis of potatoes under salt stress. *Journal of Proteome Research*, 7(11), 4858–4868.
- Agrawal, G. K., Pedreschi, R., Barkla, B. J., Bindschedler, L. V., Cramer, R., Sarkar, A., & Rakwal, R. (2012). Translational plant proteomics: A perspective. *Journal of Proteomics*, 75(15), 4588–4601.
- Aharoni, A., & Vorst, O. (2002). DNA microarrays for functional plant genomics. *Plant Molecular Biology*, 48(1), 99–118.
- Ahmad, P., Ashraf, M., Younis, M., Hu, X., Kumar, A., Akram, N. A., & Al-Qurainy, F. (2012). Role of transgenic plants in agriculture and biopharming. *Biotechnology Advances*, 30(3), 524–540.
- Akbaba, U., Sahin, Y., & Turkez, H. (2012). Comparison of element contents in haricot beans grown under organic and conventional farming regimes for human nutrition and health. *Acta Scientiarum Polonorum-Hortorum Cultus*, 11(2), 117–125.
- Ashikari, M., Sakakibara, H., Lin, S., Yamamoto, T., Takashi, T., Nishimura, A., & Matsuoka, M. (2005). Cytokinin oxidase regulates rice grain production. *Science*, 309(5735), 741–745.
- Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: technologies and their applications. *Journal of Chromatographic Science*, 55(2), 182–196.
- Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., & Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, 465(7298), 627–631.
- Babalola, O. O. (2010). Beneficial bacteria of agricultural importance. *Biotechnology Letters*, 32(11), 1559–1570.
- Bachlava, E., Taylor, C. A., Tang, S., Bowers, J. E., Mandel, J. R., Burke, J. M., & Knapp, S. J. (2012). SNP discovery and development of a high-density genotyping array for sunflowers. *PLoS One*, 7(1), e29814.
- Baker, N. R. (2008). Chlorophyll fluorescence: a probe of photosynthesis in vivo. *Annual Review of Plant Biology*, 59, 89–113.
- Barbier-Brygoo, H., & Joyard, J. (2004). Focus on plant proteomics. *Plant Physiology and Biochemistry*, 42(12), 913–917.
- Baxter, I. (2009). Ionomics: Studying the social network of mineral nutrients. *Current Opinion in Plant Biology*, 12(3), 381–386.
- Bennett, S. T., Barnes, C., Cox, A., Davies, L., & Brown, C. (2005). Toward the \$1000 human genome. *Pharmacogenomics*, 6(4), 373–382.
- Berger, B., Parent, B., & Tester, M. (2010). High-throughput shoot imaging to study drought responses. *Journal of Experimental Botany*, 61(13), 3519–3528.
- Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., & Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*, 9(9), 418–425.
- Borisjuk, L., Rolletschek, H., & Neuberger, T. (2012). Surveying the plant’s world by magnetic resonance imaging. *The Plant Journal*, 70(1), 129–146.
- Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, 12(10), 1–8.
- Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L., D’Amore, R., Allen, A. M., & Hall, N. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426), 705–710.
- Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., & Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18(6), 630–634.
- Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., & Borevitz, J. O. (2014). TraitCapture: genomic and environment modelling of plant phenomic data. *Current Opinion in Plant Biology*, 18, 73–79.
- Buggs, R. J., Renny-Byfield, S., Chester, M., Jordan-Thaden, I. E., Viccini, L. F., Chamala, S., & Soltis, D. E. (2012). Next-generation sequencing and genome evolution in allopolyploids. *American Journal of Botany*, 99(2), 372–382.
- Carreno-Quintero, N., Bouwmeester, H. J., & Keurentjes, J. J. (2013). Genetic analysis of metabolome–phenotype interactions: From model to crop species. *Trends in Genetics*, 29(1), 41–50.
- Challam, C., Nandhakumar, N., & Kardile, H. B. (2019). Advances in crop improvement: Use of miRNA technologies for crop improvement. *OMICS-Based Approaches in Plant Biotechnology*, 55–74.
- Chen, H., Xie, W., He, H., Yu, H., Chen, W., Li, J., & Zhang, Q. (2014). A high-density SNP genotyping array for rice biology and molecular breeding. *Molecular Plant*, 7(3), 541–553.
- Clevenger, J. P., Korani, W., Ozias-Akins, P., & Jackson, S. (2018). Haplotype-based genotyping in polyploids. *Frontiers in Plant Science*, 9, 564.
- Collard, B. C., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491), 557–572.
- Coopersmith, E. J., Minsker, B. S., Wenzel, C. E., & Gilmore, B. J. (2014). Machine learning assessments of soil drying for agricultural planning. *Computers and Electronics in Agriculture*, 104, 93–104.
- Coppens, F., Wuyts, N., Inzé, D., & Dhondt, S. (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Current Opinion in Systems Biology*, 4, 58–63.
- Daub, C. O., Kloska, S., & Selbig, J. (2003). MetaGeneAlyse: Analysis of integrated transcriptional and metabolite data. *Bioinformatics*, 19(17), 2332–2333.
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491.

- Dixon, R. A., Gang, D. R., Charlton, A. J., Fiehn, O., Kuiper, H. A., Reynolds, T. L., & Seiber, J. N. (2006). Applications of metabolomics in agriculture. *Journal of Agricultural and Food Chemistry*, *54*(24), 8984–8994.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., & Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.
- Farre, G., Twyman, R. M., Christou, P., Capell, T., & Zhu, C. (2015). Knowledge-driven approaches for engineering complex metabolic pathways in plants. *Current Opinion in Biotechnology*, *32*, 54–60.
- Fernie, A. R., & Schauer, N. (2009). Metabolomics-assisted breeding: A viable option for crop improvement? *Trends in Genetics*, *25*(1), 39–48.
- Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S., IV (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology*, *54*(1), 357–374.
- Galinha, C., Anawar, H. M., Freitas, M. D. C., Pacheco, A. M. G., Almeida-Silva, M., Coutinho, J., & Almeida, A. S. (2011). Neutron activation analysis of wheat samples. *Applied Radiation and Isotopes*, *69*(11), 1596–1604.
- Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., & Falque, M. (2011). A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS one*, *6*(12), e28334.
- Garcia-Cela, E., Kiaitsi, E., Medina, A., Sulyok, M., Krska, R., & Magan, N. (2018). Interacting environmental stress factors affects targeted metabolomic profiles in stored natural wheat and that inoculated with *F. graminearum*. *Toxins*, *10*(2), 56.
- Gilroy, S., & Jones, D. L. (2000). From form to function: Development and nutrient uptake in root hairs. *Trends in Plant Science*, *5*, 56–60.
- Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., & Briggs, S. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, *296*(5565), 92–100.
- Gong, C. Y., & Wang, T. (2013). Proteomic evaluation of genetically modified crops: Current status and challenges. *Frontiers in Plant Science*, *4*, 41.
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186.
- Griffiths, J. F., Griffiths, A. J., Wessler, S. R., Lewontin, R. C., Gelbart, W. M., Suzuki, D. T., & Miller, J. H. (2005). *An introduction to genetic analysis*. Macmillan.
- Hartmann, A., Czauderna, T., Hoffmann, R., Stein, N., & Schreiber, F. (2011). HTPPheno: An image analysis pipeline for high-throughput plant phenotyping. *BMC Bioinformatics*, *12*(1), 1–9.
- Heffner, E. L., Jannink, J. L., & Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome*, *4*(1).
- Herman, E. M., Helm, R. M., Jung, R., & Kinney, A. J. (2003). Genetic modification removes an immunodominant allergen from soybean. *Plant Physiology*, *132*(1), 36–43.
- Howarth, C.J., Gay, A.P., Draper, J., & Powell, W. (2011, January). Development of high throughput plant phenotyping facilities at Aberystwyth. In *Plant and Animal Genome XIX Conference*.
- Hu, H., Scheben, A., & Edwards, D. (2018). Advances in integrating genomics and bioinformatics in the plant breeding pipeline. *Agriculture*, *8*(6), 75.
- Hu, J., Rampitsch, C., & Bykova, N. V. (2015). Advances in plant proteomics toward improvement of crop productivity and stress resistance. *Frontiers in Plant Science*, *6*, 209.
- Jude Immaculate, H., Evanzalin Ebananjar, P., Sivaranjani, K., & Sebastian Terence, J. (2020). *Applications of machine learning algorithms in agriculture*.
- Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., & Somerville, C. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *nature*, *408*(6814), 796–815.
- Kersten, B., Bürkle, L., Kuhn, E. J., Giavalisco, P., Konthur, Z., Lueking, A., & Schneider, U. (2002). Large-scale plant proteomics. *Functional Genomics*, 133–141.
- Kessler, N., Neuweger, H., Bonte, A., Langenkämper, G., Niehaus, K., Nattkemper, T. W., & Goesmann, A. (2013). MeltDB 2.0—Advances of the metabolomics software system. *Bioinformatics*, *29*(19), 2452–2459.
- Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., & Gao, G. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machines. *Nucleic acids research*, *35*(suppl_2), W345–W349.
- Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., & Burt, D. W. (2013). Development of a high density 600 K SNP genotyping array for chicken. *BMC genomics*, *14*(1), 1–13.
- Kwong, Q. B., Teh, C. K., Ong, A. L., Heng, H. Y., Lee, H. L., Mohamed, M., & Appleton, D. R. (2016). Development and validation of a high-density SNP genotyping array for African oil palm. *Molecular plant*, *9*(8), 1132–1141.
- L’vov, B. V. (2005). Fifty years of atomic absorption spectrometry. *Journal of Analytical Chemistry*, *60*(4), 382–392.
- Lahner, B., Gong, J., Mahmoudian, M., Smith, E. L., Abid, K. B., Rogers, E. E., & Salt, D. E. (2003). Genomic scale profiling of nutrient and trace elements in *Arabidopsis thaliana*. *Nature biotechnology*, *21*(10), 1215–1221.
- Langridge, P., & Fleury, D. (2011). Making the most of ‘omics’ for crop breeding. *Trends in Biotechnology*, *29*(1), 33–40.
- Lepcha, P., Kumar, P. R., & Sathyanarayana, N. (2019). Exploring genomics research in the context of some underutilized legumes—A review. *OMICS-Based Approaches in Plant Biotechnology*, 1–18.
- Lister, R., Gregory, B. D., & Ecker, J. R. (2009). Next is now: New technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology*, *12*(2), 107–118.
- Liu, L., Mei, Q., Yu, Z., Sun, T., Zhang, Z., & Chen, M. (2013). An integrative bioinformatics framework for genome-scale multiple level network reconstruction of rice. *Journal of Integrative Bioinformatics*, *10*(2), 94–102.

- Li, N., Koh, Z. X., Goh, J., Lin, Z., Haaland, B., Ting, B. P., & Ong, M. E. H. (2014). Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection. *BMC Medical Informatics and Decision Making*, *14*(1), 1–9.
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), e1005457.
- Lu, Y., Savage, L. J., Larson, M. D., Wilkerson, C. G., & Last, R. L. (2011). Chloroplast 2010: A database for large-scale phenotypic screening of *Arabidopsis* mutants. *Plant physiology*, *155*(4), 1589–1600.
- Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., & Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome biology*, *16*(1), 1–21.
- Mapleson, D., Venturini, L., Kaitthakottil, G., & Swarbreck, D. (2017). Efficient and accurate detection of splice junctions from RNAseq with Portcullis. bioRxiv 217620. *Doi*, *10*, 217620.
- Marschner, H. (2011). *Marschner's mineral nutrition of higher plants*. Academic Press.
- Marx, H., Minogue, C. E., Jayaraman, D., Richards, A. L., Kwiecien, N. W., Siahpirani, A. F., & Coon, J. J. (2016). A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. *Nature biotechnology*, *34*(11), 1198.
- Maxwell, K., & Johnson, G. N. (2000). Chlorophyll fluorescence—A practical guide. *Journal of experimental botany*, *51*(345), 659–668.
- McCouch, S. R., Zhao, K., Wright, M., Tung, C. W., Ebana, K., Thomson, M., & Bustamante, C. (2010). Development of genome-wide SNP assays for rice. *Breeding Science*, *60*(5), 524–535.
- McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology*, *17*(1), 4–11.
- Mehrotra, S., & Goyal, V. (2013). Evaluation of designer crops for biosafety—A scientist's perspective. *Gene*, *515*(2), 241–248.
- Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature reviews genetics*, *11*(1), 31–46.
- Michael, T. P., & VanBuren, R. (2015). Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology*, *24*, 71–81.
- Mishra, J., & Arora, N. K. (2018). Secondary metabolites of fluorescent pseudomonads in biocontrol of phytopathogens for sustainable agriculture. *Applied Soil Ecology*, *125*, 35–45.
- Miyao, A., Iwasaki, Y., Kitano, H., Itoh, J. I., Maekawa, M., Murata, K., & Hirochika, H. (2007). A large-scale collection of phenotypic data describing an insertional mutant population to facilitate functional analysis of rice genes. *Plant Molecular Biology*, *63*(5), 625–635.
- Mochida, K., & Shinozaki, K. (2010). Genomics and bioinformatics resources for crop improvement. *Plant and Cell Physiology*, *51*(4), 497–523.
- Montenegro-Burke, J. R., Aisporna, A. E., Benton, H. P., Rinehart, D., Fang, M., Huan, T., & Siuzdak, G. (2017). Data streaming for metabolomics: accelerating data processing and analysis from days to minutes. *Analytical Chemistry*, *89*(2), 1254–1259.
- Moose, S. P., & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology*, *147*(3), 969–977.
- Morellos, A., Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R., Tziotziou, G., & Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, 104–116.
- Munns, R., James, R. A., Sirault, X. R., Furbank, R. T., & Jones, H. G. (2010). New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. *Journal of Experimental Botany*, *61*(13), 3499–3507.
- Naik, H. S., Zhang, J., Lofquist, A., Assefa, T., Sarkar, S., Ackerman, D., & Ganapathysubramanian, B. (2017). A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant Methods*, *13*(1), 1–12.
- Nasarudin, N. E. M., & Shafri, H. Z. M. (2011). Development and utilization of urban spectral library for remote sensing of urban environment. *Journal of Urban and Environmental Engineering*, *5*(1), 44–56.
- Natarajan, S., Xu, C., Bae, H., & Bailey, B. A. (2007). Proteomic and genomic characterization of Kunitz trypsin inhibitors in wild and cultivated soybean genotypes. *Journal of Plant Physiology*, *164*(6), 756–763.
- Oksman-Caldentey, K. M., & Saito, K. (2005). Integrating genomics and metabolomics for engineering plant metabolic pathways. *Current Opinion in Biotechnology*, *16*(2), 174–179.
- Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98.
- Pandit, A. A., Shah, R. A., & Husaini, A. M. (2018). Transcriptomics: A time-efficient tool with wide applications in crop and animal biotechnology. *Journal of Pharmacognosy and Phytochemistry*, *7*(2), 1701–1704.
- Pasala, R., & Pandey, B. B. (2020). Plant phenomics: High-throughput technology for accelerating genomics. *Journal of Biosciences*, *45*(1), 1–6.
- Patrick, J. W., Botha, F. C., & Birch, R. G. (2013). Metabolic engineering of sugars and simple sugar derivatives in plants. *Plant Biotechnology Journal*, *11*(2), 142–156.
- Pertea, M. (2012). The human transcriptome: An unfinished story. *Genes*, *3*(3), 344–360.
- Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T., & Brazma, A. (2016). Expression Atlas update—An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, *44*(D1), D746–D752.
- Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., & Jannink, J. L. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome*, *5*(3).
- Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human Molecular Genetics*, *27*(R2), R234–R241.
- Poole, R., Barker, G., Wilson, I. D., Coghill, J. A., & Edwards, K. J. (2007). Measuring global gene expression in polyploidy; a cautionary note from allohexaploid wheat. *Functional and Integrative Genomics*, *7*(3), 207–219.
- Rafalski, J. A. (2002). Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science*, *162*(3), 329–333.
- Rahman, H., Ramanathan, V., Jagadeeshselvam, N., Ramasamy, S., Rajendran, S., Ramachandran, M., & Muthurajan, R. (2015). *Phenomics: Technologies and applications in plant and agriculture*. *PlantOmics: The omics of plant science* (pp. 385–411). New Delhi: Springer.

- Razzaq, A., Sadia, B., Raza, A., Khalid Hameed, M., & Saleem, F. (2019). Metabolomics: A way forward for crop improvement. *Metabolites*, 9(12), 303.
- Rhee, S. Y., Dickerson, J., & Xu, D. (2006). Bioinformatics and its applications in plant biology. *Annual Review of Plant Biology*, 57, 335–360.
- Rich-Griffin, C., Stechemesser, A., Finch, J., Lucas, E., Ott, S., & Schäfer, P. (2020). Single-cell transcriptomics: A high-resolution avenue for plant functional genomics. *Trends in Plant Science*, 25(2), 186–197.
- Rincon, G., Weber, K. L., Van Eenennaam, A. L., Golden, B. L., & Medrano, J. F. (2011). Hot topic: Performance of bovine high-density genotyping platforms in Holsteins and Jerseys. *Journal of Dairy Science*, 94(12), 6116–6121.
- Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., & Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), 348–352.
- Ruan, Y., Le Ber, P., Ng, H. H., & Liu, E. T. (2004). Interrogating the transcriptome. *Trends in Biotechnology*, 22(1), 23–30.
- Saha, G. C., Sarker, A., Chen, W., Vandemark, G. J., & Muehlbauer, F. J. (2010). Inheritance and linkage map positions of genes conferring resistance to *Stemphylium* blight in lentils. *Crop Science*, 50(5), 1831–1839.
- Salgotra, R. K., Gupta, B. B., & Stewart, C. N. (2014). From genomics to functional markers in the era of next-generation sequencing. *Biotechnology Letters*, 36(3), 417–426.
- Salt, D. E., Baxter, I., & Lahner, B. (2008). Ionomics and the study of the plant ionome. *Annual Review of Plant Biology*, 59, 709–733.
- Saxena, R. K., Rathore, A., Bohra, A., Yadav, P., Das, R. R., Khan, A. W., & Varshney, R. K. (2018). Development and application of high-density Axiom *Cajanus* SNP array with 56 K SNPs to understand the genome architecture of released cultivars and founder genotypes. *The Plant Genome*, 11(3), 180005.
- Schenk, P. M., Carvalhais, L. C., & Kazan, K. (2012). Unraveling plant–microbe interactions: can multi-species transcriptomics help? *Trends in Biotechnology*, 30(3), 177–184.
- Semba, R. D. (2016). The rise and fall of protein malnutrition in global health. *Annals of Nutrition and Metabolism*, 69(2), 79–88.
- Severin, A. J., Woody, J. L., Bolon, Y. T., Joseph, B., Diers, B. W., Farmer, A. D., & Shoemaker, R. C. (2010). RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome. *BMC Plant Biology*, 10(1), 1–16.
- Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, 21(2), 110–124.
- Singh, N., Bhatt, V., Rana, N., & Shivaraj, S. M. (2020). *Advances of next-generation sequencing (NGS) technologies to enhance the biofortification in crops*. *Advances in Agri-Food Biotechnology* (pp. 427–450). Singapore: Springer.
- Singh, N., Jayaswal, P. K., Panda, K., Mandal, P., Kumar, V., Singh, B., & Singh, N. K. (2015). Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. *Scientific Reports*, 5(1), 1–9.
- Singh, N., Rai, V., & Singh, N. K. (2020). Multi-omics strategies and prospects to enhance seed quality and nutritional traits in pigeonpea. *The Nucleus*, 1–8.
- Singh, S., Mahato, A. K., Jayaswal, P. K., Singh, N., Dheer, M., Goel, P., & Singh, N. K. (2020). A 62 K genic-SNP chip array for genetic studies and breeding applications in pigeonpea (*Cajanus cajan* L. Millsp. *Scientific Reports*, 10(1), 1–14.
- Slaughter, D. C., Giles, D. K., & Downey, D. (2008). Autonomous robotic weed control systems: A review. *Computers and Electronics in Agriculture*, 61(1), 63–78.
- Smit, Z. (2005). Recent developments of material analysis with PIXE. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 240(1–2), 258–264.
- Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., & Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring*, 27(6), 747–751.
- Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One*, 8(1), e54985.
- Souda, P., Ryan, C. M., Cramer, W. A., & Whitelegge, J. (2011). Profiling of integral membrane proteins and their post translational modifications using high-resolution mass spectrometry. *Methods*, 55(4), 330–336.
- Sozzani, R., Busch, W., Spalding, E. P., & Benfey, P. N. (2014). Advanced imaging techniques for the study of plant growth and development. *Trends in Plant Science*, 19(5), 304–310.
- Talaviya, T., Shah, D., Patel, N., Yagnik, H., & Shah, M. (2020). Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artificial Intelligence in Agriculture*.
- Tello-Ruiz, M. K., Stein, J., Wei, S., Preece, J., Olson, A., Naithani, S., & Ware, D. (2016). Gramene 2016: Comparative plant genomics and pathway resources. *Nucleic Acids Research*, 44(D1), D1133–D1140.
- Tsaftaris, S. A., & Noutsos, C. (2009). *Plant phenotyping with low cost digital cameras and image analytics*. *Information Technologies in Environmental Engineering* (pp. 238–251). Berlin, Heidelberg: Springer.
- Tulchinsky, T. H. (2010). Micronutrient deficiency conditions: Global health issues. *Public Health Reviews*, 32(1), 243–255.
- United Nations Food and Agriculture Organization. Dimensions of need - An atlas of food and agriculture. Staple foods: What do people eat [Internet]. 2015. Available from: <http://www.fao.org/docrep/u8480e/u8480e07.htm> [Accessed: 2015-11-10].
- Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., & Schön, C. C. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*, 15(1), 1–15.
- Urano, K., Maruyama, K., Ogata, Y., Morishita, Y., Takeda, M., Sakurai, N., & Shinozaki, K. (2009). Characterization of the ABA-regulated global responses to dehydration in *Arabidopsis* by metabolomics. *The Plant Journal*, 57(6), 1065–1078.

- Valdés, A., Ibáñez, C., Simó, C., & García-Cañas, V. (2013). Recent transcriptomics advances and emerging applications in food science. *TrAC Trends in Analytical Chemistry*, 52, 142–154.
- Van Borm, S., Belák, S., Freimanis, G., Fusaro, A., Granberg, F., Höper, D., & Rosseel, T. (2015). *Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases? Veterinary infection biology: molecular diagnostics and high-throughput strategies* (pp. 415–436).
- Van der Vlugt, R., Minafra, A., Olmos, A., Ravnkar, M., Wetzel, T., Varveri, C., & Massart, S. (2015). *Application of next generation sequencing for study and diagnosis of plant viral diseases in agriculture*.
- Van Emon, J. M. (2016). The omics revolution in agricultural research. *Journal of Agricultural and Food Chemistry*, 64(1), 36–44.
- Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., & Cook, D. R. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology*, 31(3), 240–246.
- Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2000). Analysing uncharted transcriptomes with SAGE. *Trends in Genetics*, 16(10), 423–425.
- Villas-Boas, S. G., Koulman, A., & Lane, G. A. (2007). Analytical methods from the perspective of method standardization. *Metabolomics*, 11–52.
- Walley, J. W., Shen, Z., McReynolds, M. R., Schmelz, E. A., & Briggs, S. P. (2018). Fungal-induced protein hyperacetylation in maize identified by acetylome profiling. *Proceedings of the National Academy of Sciences*, 115(1), 210–215.
- Wang, H., Cimen, E., Singh, N., & Buckler, E. (2020). Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology*, 4(54), 34–41. Available from <https://doi.org/10.1016/j.pbi.2019.120.010>.
- Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., & Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal*, 12(6), 787–796.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57–63.
- White, P. J., & Broadley, M. R. (2009). Biofortification of crops with seven mineral elements often lacking in human diets—iron, zinc, copper, calcium, magnesium, selenium and iodine. *New Phytologist*, 182(1), 49–84.
- Wilson, S. A., & Roberts, S. C. (2014). Metabolic engineering approaches for production of biochemicals in food and medicinal plants. *Current Opinion in Biotechnology*, 26, 174–182.
- Winfield, M. O., Allen, A. M., Burrige, A. J., Barker, G. L., Benbow, H. R., Wilkinson, P. A., & Edwards, K. J. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnology Journal*, 14(5), 1195–1206.
- Wishart, D. S. (2011). Advances in metabolite identification. *Bioanalysis*, 3(15), 1769–1782.
- Xu, C., Xu, Y., & Huang, B. (2008). Protein extraction for two-dimensional gel electrophoresis of proteomic profiling in turfgrass. *Crop Science*, 48(4), 1608–1614.
- Yadav, P.K., Kumar, S., Kumar, S., & Yadav, R.C. (2018). *Crop improvement for sustainability*.
- Yang, L., Fountain, J. C., Ji, P., Ni, X., Chen, S., Lee, R. D., & Guo, B. (2018). Deciphering drought-induced metabolic responses and regulation in developing maize kernels. *Plant Biotechnology Journal*, 16(9), 1616–1628.
- Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., & Yan, J. (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Molecular Plant*, 13(2), 187–214.
- Yilmaz, A., Nishiyama, M. Y., Fuentes, B. G., Souza, G. M., Janies, D., Gray, J., & Grotewold, E. (2009). GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiology*, 149(1), 171–180.
- Young, N. D., Debellé, F., Oldroyd, G. E., Geurts, R., Cannon, S. B., Udvardi, M. K., & Roe, B. A. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378), 520–524.
- Yu, P., & Lin, W. (2016). Single-cell transcriptome study as big data. *Genomics, Proteomics and Bioinformatics*, 14(1), 21–30.
- Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2017). Improvements in genomic technologies: Application to crop genomics. *Trends in Biotechnology*, 35(6), 547–558.
- Zhang, T., Zhang, X., Hu, S., & Yu, J. (2011). An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. *Plant Methods*, 7(1), 1–8.
- Zhao, C., Zhang, Y., Du, J., Guo, X., Wen, W., Gu, S., & Fan, J. (2019). Crop phenomics: Current status and perspectives. *Frontiers in Plant Science*, 10, 714.
- Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., & McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature communications*, 2(1), 1–10.
- Zhao, Q., Chen, S., & Dai, S. (2013). C4 photosynthetic machinery: insights from maize chloroplast proteomics. *Frontiers in Plant Science*, 4, 85.
- Zhen, Y., Qi, J. L., Wang, S. S., Su, J., Xu, G. H., Zhang, M. S., & Yang, Y. H. (2007). Comparative proteome analysis of differentially expressed proteins induced by Al toxicity in soybean. *Physiologia Plantarum*, 131(4), 542–554.
- Zivy, M., Wienkoop, S., Renaut, J., Pinheiro, C., Goulas, E., & Carpentier, S. (2015). The quest for tolerant varieties: the importance of integrating “omics” techniques to phenotyping. *Frontiers in Plant Science*, 6, 448.

This page intentionally left blank

Chapter 2

Promises and benefits of omics approaches to data-driven science industries

Niranjani Iyer

Biovia Corp, Dassault Systemes, San Diego, CA, United States

Globally, 8 billion people have to be fed, and this number is expected to reach 10 billion by the year 2050. The UN has a sustainable goal to eliminate hunger by 2030; nearly 690 million people worldwide still are unable get a single meal a day. There are three important challenges to human and planetary health and that include feeding a growing and increasing population, reducing environmental impact, and growing plants that can adapt to changing climate conditions. The One Health initiative aims to work locally, nationally, and globally to achieve optimal health for people, animals, and the planet (https://en.wikipedia.org/wiki/One_Health). In order to achieve this lofty initiative, a holistic approach needs to be adopted and implemented. An example of holistic approach would be deciphering the genomes of the organisms of interest in a given environment. Thus sequencing a plant or a human genome is not sufficient by itself. The soil rhizosphere in which the plant is growing and the microbes on the plant are important. On the same grounds, the gut microbiome from the humans is vital to gain a holistic understanding of the human genome. Genomic resources expand the toolbox available for plant breeding and crop improvement efforts. This chapter will focus on various types of omics approach, and data-driven science in revolutionizing genomic science, and helping human health to agricultural science.

In the last two decades, science is trending to be more multiscale high-throughput data. With the advancements in sequencing technologies, large-scale omics analysis has revolutionized biology. Our understanding of biological processes is largely driven by omics datasets that includes genomics, transcriptomics, proteomics, epigenomics, metabolomics, etc. With the decreasing cost of sequencing, rapid generation of data creates exciting opportunities and formidable challenges. This chapter will discuss about various sequencing technologies, different omics analysis, and the challenges of data integration. The use of machine learning and AI as potential tools for studying the vast multiplatform omics datasets is discussed.

2.1 Sequencing technologies

The building block of life is the treasure trove that contains genetic information carried in the DNA of the organism. Scientists believe that sequencing genomes and understanding the coding regions and the noncoding regions that carry out regulatory instructions can highlight many of the biology of interest. The challenge of growing population and climate changes, enhancing genetic gain in biotechnology using multipronged approach and combining conventional and genomic technologies holds potential promise for biotechnology industries.

Within crop genomics, advances relevant to crop improvement have primarily been in marker [e.g., Illumina single-nucleotide polymorphism (SNP) chips, Kompetitive allele-specific PCR (KASP) assays, genotyping by sequencing, and sequencing (e.g., Illumina, PacBio, Nanopore)] technology. Recent innovations are driving a paradigm shift in which the extent and relevance of structural variation within the pan-genome of crop species are now being considered (Coletta et al., 2021).

Early 2000 was the first step in the initial exploration to study plant and animal genomes using the next-generation sequencing (NGS) technologies ([Arabidopsis Genome Initiative, 2000](#); [International Rice Genome Sequencing Project, 2005](#); [Schnable et al., 2009](#)). Several sequencing projects were initiated, and genome assemblies of many plants were completed. These studies revolutionized the way biology and breeding of crops were done. Early plant genome assemblies revealed interesting diversity in plant genome. This was observed at SNP level and structural variants (SVs) (e.g., presence–absence variation and copy number variation), and chromosomal rearrangements, and repetitive portions of the genome [e.g., transposable elements (TEs), knob repeats, and centromere repeats]. All these interesting features helped in characterizing the “core genome,” that is, common to all organisms within a species and the rest as “dispensable” genome.

2.2 Advances in genome assembly technology

The advent of NGS technologies in the last two decades and assembly algorithms profoundly helped in understanding the complexity of the genomes. NGS enables whole-genome sequencing (WGS), and resequencing, transcriptome sequencing, metagenomics, and high-throughput genotyping. These techniques can be applied to understand genetic diversity, genetic and epigenetic characterization of genomes.

Sequencing of genes in late 1970s was based on Sanger sequencing. This method was expensive and was adopted for genome sequencing later by doing BAC libraries [(e.g., rice ([International Rice Genome Sequencing Project, 2005](#)), maize ([Schnable et al., 2009](#)), sorghum ([Paterson et al., 2009](#)), and soybean ([Schmutz et al., 2010](#))]. When NGS technology came into picture, crop reference genomes were done using paired-end and mate-pair Illumina data and de Bruijn graph approaches [e.g., barley ([Schmutz et al., 2010](#)) and wheat ([International Wheat Genome Sequencing Consortium IWGSC et al., 2018](#))]. The reduced cost of Illumina sequencing and improved assembly algorithms facilitated accession level genome assembly. Several de novo assembly techniques were adopted to build multiple accessions per crop using low-cost short-read data [e.g., maize-PH207 ([Hirsch et al., 2016](#)), maize-W22 ([Springer et al., 2018](#)); maize-HZS ([Li et al., 2019](#)), maize-Flint genomes ([Haberer et al., 2020](#)), rice genomes ([Schatz et al., 2014](#)), and soybean genomes ([Li, Fillmore, et al., 2014](#); [Li, Zhou, et al., 2014](#))].

Sequencing by short reads approach was extremely economical but genome assembly was a challenge in repetitive, TE-rich regions of the genomes and the regions closer to the centromere. This resulted in several draft genomes with numerous gaps and partial assembly. Often Sanger sequencing would be adopted to close the gaps or assemble the draft genome. PacBio offered the long-read technology that facilitated addressing some of the shortcomings of the small-read techniques. Although this technique led to discoveries of variation and copy number ([Song et al., 2020](#); [Zhou, Hirsch, Briggs, & Springer, 2019](#)), it had a high error rate in base calling. Improvements in this technology have considerably reduced the error rates and facilitated long-read assemblies for uncovering agronomically relevant information across different lines within crop species. It is important to understand the different sequencing technologies and data analysis steps in assembling the genome.

2.2.1 Algorithms in reference-based and de novo assembly

Sequencing technologies for short reads and long reads of WGS provide the information of entire genetic material of an organism. There are two main approaches involved in assembling these reads into longer contiguous genomic sequences. Both of these methods have their pros and cons and its often-scientific subject knowledge drives the decision on which method would work best for their type of studies.

Prior to assembly, quality checks of raw reads are critical in determining the output of the assembly. High-quality reads are important for downstream assembly algorithms and analysis. The raw reads are subjected to quality filtration and adapter trimming. One of the popular algorithms used is Trimmomatic ([Bolger, Lohse, & Usadel, 2014](#)). The primer sequences, polyA tails, and reads produced from ribosomal DNA template are trimmed out.

The reference-based assembly approach involves mapping each read to a reference genome sequence to identify genetic variation such as SNPs, indels, insertions, copy number variants, genome-wide association studies (GWAS), and building haplotypes from genome assemblies. The reference-based assembly is usually done by downloading the reference sequence (fasta) and genome and gene information (GFF3 or GTF) from public databases and indexing them. The high-quality fastq files are aligned with the reference genome using BWA mem (Bowtie, BWA are other commonly used algorithms) with optimized parameters. The reads are aligned to the reference index.

When no reference sequences are available, a de novo-based approach is used, where sequenced reads are compared to each other, and then overlapped reads are used to build longer contiguous sequences. The built contigs are then

oriented. The de novo assembly of high-quality reads is usually used with some popular de novo algorithms such as Velvet/Spades/SOAP de novo. These assemblers are highly sensitive to input parameters and multiple Kmer assembly runs are done to optimize the assembly. This is assembled using Kmer length, coverage cutoff, insert length, and expected coverage for scaffold assembly. The best assembly is usually selected based on scaffold N50 and max scaffold length. Given the variability observed in de novo techniques, the final assembly is determined based on scaffolds N50, assembly coverage (depth), reads participated in assembly, and guanine-cytosine (GC) content assembly.

2.2.2 Postassembly algorithms for encoding the biology

Once a genome is assembled by either reference or de novo based, several steps go in to understand the biology of the genome. The first step includes the gene prediction. Ab initio, gene predictions are statistical models and are trained to find features of genes start and stop codons, and alternate splicing. Gene prediction algorithm such as Prodigal/Augustus is popular in predicting the coding regions in the draft genome. The next layer of biology is the annotation of the coding regions. Often this involves bringing in all the knowledge from several databases such as NCBI nonredundant protein database (NR), Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), Cluster of Orthologous Group (COG), Pfam, and Gene Ontology to classify the genes based on biological processes, molecular function, and cellular localization.

One of the major advancement with genomic assembly technology is often interesting features in the genomic studies such as SNPs and Indel and structural variations are valuable tools to associate with a biological process. Many a times, SNP discovery is done from the alignment file generated by BWA/Bowtie program. SAMTOOLS/GATK tools with optimized parameters are used in calling SNPs and Indels. Comparative genome analysis or synteny analysis is done with closely related species by pairwise alignments or OrthoMCL. These methods are also often used in identification of core genes in a species.

The de novo assembly of genomes results in a nonreference-based manner assemblies without any reference bias and often used in identification of structural variation (SV). However, this can result in number of challenges including cost, detection of false structural variants, and compromise on downstream analysis as de novo assemblies result in different output depending on the data types, algorithms parameters, and the assembly algorithms. The other major challenge can be the consolidation of pan-genome variation into a single reference system that can affect the biological significance of SV in QTL analysis, GWAS, and genomic prediction. In pan-genome context, SV is called by mapping resequencing reads to a reference.

Several methods exist for SV information in a pan-genome context. One approach known as map-to-pan approach is to map resequencing reads to a reference genome, de novo assemble unmapped reads, and add the assembled contigs to the reference assembly (Golicz et al., 2016; Hu, Wei, & Li, 2020). This strategy can minimize errors by exploiting the information already available from a high-quality reference genome and limit the coordinate consolidation issue. Yet the genomic locations of newly assembled contigs remain unknown without further analysis. An alternative approach is the construction of a graph-based rather than linear reference genome (Computational Pan-Genomics Consortium). In this method, any variant such as SNP or SV are added to the reference as a node at the genomic location where it is discovered (Garrison et al., 2018; Rakocevic et al., 2019). Based on the strengths of the graph and linear method, recently a hybrid approach was developed. In the hybrid approach, reads are first mapped to a graph-based genome, and then haplotypes are associated with one of the reference genomes used to build the graph. Reads are then realigned to this genome and this leads to more accurate mapping than the graph-based approach alone (Grytten, Rand, Nederbragt, & Sandve, 2020).

Another important feature that was not given enough attention until recently is the widely prevalent TEs. Plant genomes are rich in TEs (Elliott & Gregory, 2015) and difficult to characterize due to repetitive fraction of genome, often creating a challenge with mapping reads to the regions. Methods to characterize variation in TE content using short-read data (Nelson, Linheiro, & Bergman, 2017) and whole-genome comparisons (Anderson et al., 2019) are emerging and will help provide access to a new level of functional variation underlying agronomic phenotypes. TEs are functionally relevant, including modifying the structure and amount of gene product that is transcribed (Alonge et al., 2020; Jiang, Bao, Zhang, Eddy, & Wessler, 2004). This is well observed in many studies. In maize, a Harbinger-like DNA transposon represses the expression of the ZmCCT9 gene to promote flowering under long-day conditions (Huang et al., 2018). In rice, a gypsy retrotransposon has been identified to enhance the expression of the OsFRDL4 gene and promote aluminum tolerance (Yokosho, Yamaji, Fujii-Kashino, & Ma, 2016). Annotating TE sequences are still a challenge and often homology-based using existing TE databases such as Repbase (Bao, Kojima, & Kohany, 2015) and P-MITE (Chen, Hu, Zhang, Lu, & Kuang, 2014).

2.2.3 Genome-wide association, a valuable tool mapping associations with a phenotype

Early crop reference genome assemblies facilitated the development of platforms (e.g., Illumina SNP chips) that allow for rapid, cost-effective genotyping of thousands or millions of SNPs across large sets of individuals. These platforms facilitated an increase in marker density which aided in the identification and cloning of QTLs associated with different traits (Kumar et al., 2017). Without extensive phenotyping, now these markers can be rapidly used for screening large populations as functional markers (Liu, He, Appels, & Xia, 2012) or through marker-assisted selection (Collard & Mackill, 2008). Sequencing assembly methods are very important for GWAS and other marker-based studies. Reference-based assemblies with single reference genomes sometimes results in reference bias, as variants associated with a trait may not be identified if it is missing in the reference genome. This has been identified when more and more accessions are sequenced, the genomes can be missing a particular gene. Certain genes are identified when reference genome assembly is different. For example, in maize, gene conferring resistance to sugarcane mosaic virus could be identified by GWAS using markers based on the B73, but not the PH207 (Gage et al., 2019). Another problem is that deletions relative to reference genomes can be misinterpreted as missing data. Although genome assembly with different algorithms can be providing slightly different results, but still these techniques have allowed us to uncover numerous new biological insights.

Although genomics provides information at the DNA level, many a times the real biological significance is to look into regions especially the gene and the transcript. Transcriptomics is the study of transcriptome, the complete set of RNA transcripts produced by the genome at any one time and how they affected by development, disease, or other environmental factors. The next section will focus on the transcriptomics and its relevance.

2.3 Transcriptomics—where genome connects to gene function

The flow of genetic information from DNA, transcribed to RNA, and then translated to protein is the central dogma of molecular biology. The study of RNA content and sum total of RNA transcripts is called “transcriptomics.” The coding region is mRNA and is the transient intermediary molecule representing the protein, while noncoding RNA (ncRNA) does not code for any proteins but perform diverse functions. The transcriptome analysis studies the set of RNA transcripts that are produced under specific conditions in a specific cell or tissue or organ. Transcriptomics have been applied to various aspects of research and field, clinical applications ranging from diagnostics and therapeutics, gene therapy applications, pharmacogenomics and disease prevention to developmental biology, evolutionary genomics, and comparative genomics.

Transcriptomic study is most commonly used to compare pairs of sample, which could be environmental conditions (abiotic or biotic stress conditions) and in developmental stages and progression of diseases or any particular state. This type of analysis provides immense datasets and often used in biomarker studies and in outcome prediction and targets for treatment. The transcriptomic analysis has a broad approach and is the most popular omics study done.

2.3.1 Methodologies and algorithms

The early study of whole transcriptome was done using microarray technology where the defined sequences (probes) were arranged on a solid substrate. The sample of total RNA was laid on the surface and the amount of binding to the probe determined the quantity of expressed genes that supposedly reflect the translation into proteins.

With the advent of NGS technology, a high-throughput RNA sequencing called as RNA-seq methods provide abundance of data information with very little starting material. In this methodology, the bulk RNA is extracted from the sample and is copied into double-stranded cDNA. The sequencing is done on any of various sequencing platform and the reads are mapped to the reference genome available in public data banks. The nucleotide sequences generated are typically around 100 bp in length but can range from 30 bp to over 10,000 bp, depending on the sequencing method used. RNA-seq leverages deep sampling of the transcriptome with many short fragments from a transcriptome to allow computational reconstruction of the original RNA transcript by aligning reads to a reference genome or to each other (de novo assembly) The expressed gene can be used in determining alternate splicing, novel transcripts, and gene fusions. The RNA-seq is not limited to genomes with reference; it can be done to study gene expression of poorly characterized species with limited genome resources. RNA-seq can be performed using many NGS platforms; however, each platform has its own requirements of sample preparation and the instrument design.

2.3.1.1 RNA-seq data analysis

One of the most popular NGS techniques involves RNA-seq analysis. These experiments generate a large volume (in millions) of raw sequence reads, which have to be processed to yield useful information. Many data analysis tools are available depending on the experimental design and goals. This analysis can be broken down into the following four stages: quality control, alignment, quantification, and differential expression (Van Verk, Hickman, Pieterse, & Van Wees, 2013). Most of these tools work in Linux environment and command-line tools can be used in servers that are in house or in cloud environment. Several R/Bioconductor and python packages are available for statistical analysis (Huber et al., 2015).

2.3.1.1.1 Quality control

The raw reads generated by a sequencing instrument are never perfect. Therefore it is important for a quality control and checking the accuracy of each base for downstream analyses. The typical QC analyses involve examining high-quality scores for base calls, GC content matches with the expected distribution, Kmers to check the overrepresentation of particularly short sequence motifs, and any unexpected high read duplication rate (Conesa et al., 2016). Some of the popular QC packages include the FastQC and FaQCs software packages (Lo & Chain, 2014). Any abnormalities identified in this step may be removed by trimming or tagging.

2.3.1.1.2 Alignment

To estimate the abundance of expression of particular gene and its spliced variants, the raw reads need to be first assembled by aligning to a reference genome or a by de novo method when the reference is not available. This comes with few challenges especially when it comes to complex genomes and technical aspects of high-performance computing. Each alignment software can provide meaningful information and unique strength in terms of speed of alignment of the short sequences, handling intron splicing, and ability to map to multiple locations. Several advancements have been addressed to increase the sequencing read length and reducing multimapping reads. A list of currently available high-throughput sequence aligners is maintained by the EBI (Fonseca, Rung, Brazma, & Marioni, 2012). Fig. 2.1 (http://cracs.fc.up.pt/~nf/hts_mappers/) lists the up-to-date compendium of high-throughput sequencing (HTS) mappers.

The DNA mappers are in blue, RNA mappers in red, miRNA mappers in green, and the bi-sulfite mappers are in purple. For more details on the sequencing platform, minimum read length, maximum read length, maximum number of mismatches and indels, and several other information can be obtained from this link (http://cracs.fc.up.pt/~nf/hts_mappers/).

Similar to the genome assembly, reference-based and de novo assemblies have their pros and cons when it comes to assembly for RNA-seq experiments. One of the main challenges with de novo assembly is the intense computational requirements compared to reference-based assembly. In addition, de novo mapping needs to be validated for gene variants. Some of the metrics used to understand the assembly of transcripts include N50 and which in some cases can be misleading and many evaluation methods have been available (Li, Fillmore, et al., 2014; Li, Zhou, et al., 2014; Smith-Unna, Bournnell, Patro, Hibberd, & Kelly, 2016). For assessment of assembly completeness, annotation-based metrics such as contig reciprocal best-hit count provide useful information.

2.3.1.1.3 Quantification

The amount of expression can be done at gene level, exon or at transcript level (spliced variants). The mapping of the annotation is done with the GFF file (general format file) or GTF file (general transfer format). The HTseq package is used in gene and exon read counts (Anders, Pyl, & Huber, 2015). For estimation of isoform, abundance from short reads is more complicated and requires probabilistic methods. Tophat cufflinks software works out to be a popular choice (Trapnell et al., 2010). Since some reads can align either equally well at multiple places, it is important to clean by removing them or align it to the most probable location. Methods like kallisto can circumvent the need of an exact alignment by using pseudo alignment and have the advantage of running faster compared to tophat/cufflink method (Bray, Pimentel, Melsted, & Pachter, 2016).

2.3.1.1.4 Differential expression

Since RNA-seq experiments are usually carried out comparing a control versus treatment sample, differential gene expression is measured by normalizing, clustering, and statistically analyzing the data. Some of the popular software

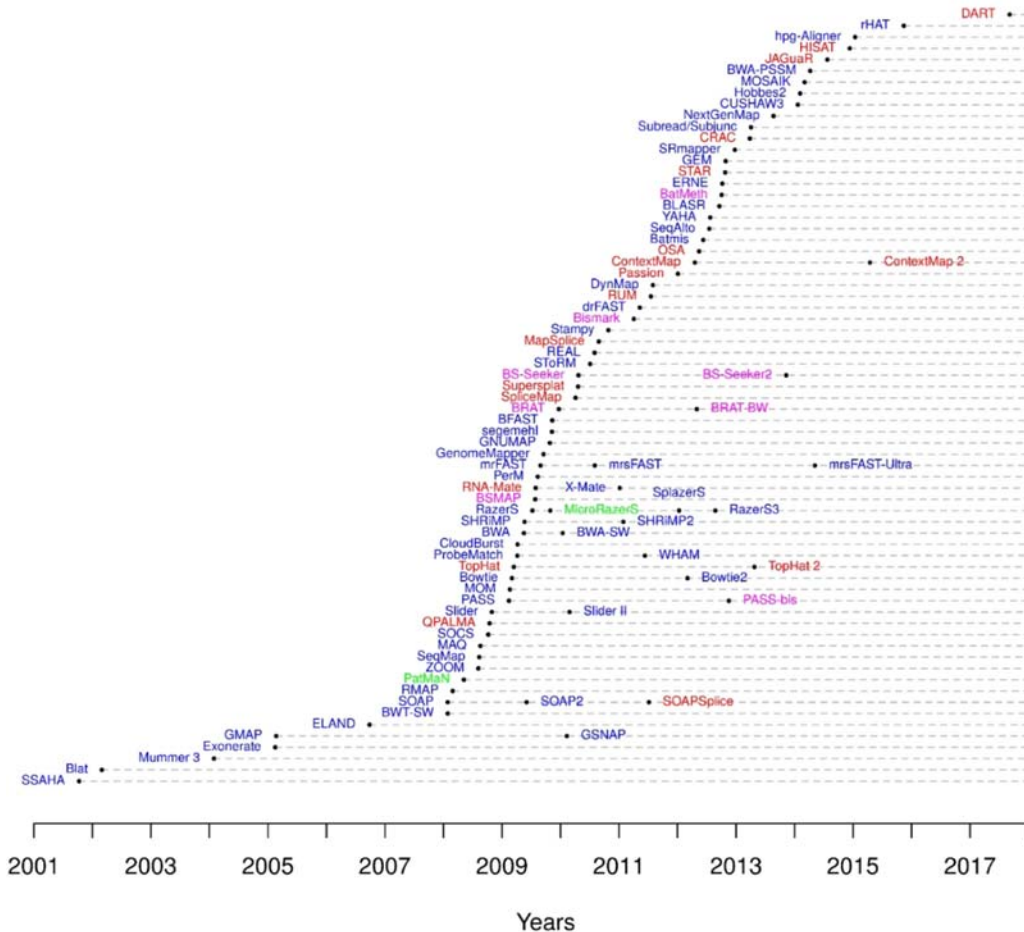


FIGURE 2.1 Compendium of HTS mapper (http://cracs.fc.up.pt/~nf/hts_mappers/).

used are the Cuffdiff2 (Linux-based) and several R/Bioconductor packages such as EdgeR, DEseq2, and Limma/Voom. Most of them read a table of genes and read counts as input, but cuffdiff uses the input bam files. The pairwise tests are common statistical tools applied for determining the differentially expressed genes or exons or isoforms.

2.3.1.2 Validating RNA-seq experiments

A standard technique that is used in validating genes of interest and their differential expression is by quantitative PCR (qPCR). This is done by measuring the expression of gene(s) of interest and control gene(s) in control and treated conditions (Fang & Cui, 2011). This method is restricted to smaller regions of less than 300 bp and targeting more to the 3'-end of the coding region. Regions in isoforms that determine the differences are targeted for primer design for discrimination of the different isoforms or spliced variants. This technique has been considered as a validation test and shows strong correlation to RNA-seq data (qPCR validation of RNA-seq data has generally shown that different RNA-seq methods are highly correlated (Camarena, Bruno, Euskirchen, Poggio, & Snyder, 2010; Core, Waterfall, & Lis, 2008).

2.3.2 Noncoding RNA

The data analysis techniques that are applied for mRNA is applied for the noncoding regions in the genome. Since the functions of the noncoding regions are associated with transcriptional regulation, RNA splicing, DNA replication etc., often the biological insights from these experiments are done using the databases for small RNAs (Table 2.1).

TABLE 2.1 List of some databases specially designed for transcriptome analysis.

Database	Host	Description
Gene Expression Omnibus (GEO)	NCBI	GEO is data repository supporting MIAME-compliant data submissions for both array- and sequence-based datasets
BioStudies (previously known as Array Express)	EBI	Biostudies database offers one stop shop for all data supporting life sciences including the array data from Array Express
Expression Atlas	EBI	Public repository of gene expression pattern data under different biological conditions. This includes baseline and differential expression experiments
Genevestigator	Privately operated	This is largest preanalyzed gene expression databases in the world covering more than 250,000 microarray and RNA-seq datasets. This provides data on biomedical and plant biology, with sophisticated tools for data search, visualization, and analysis
NONCODE	Noncode.org	This database offers integrated knowledge database dedicated to noncoding RNAs (excluding tRNAs and rRNAs)

2.3.3 Epigenomics

Only 1% of the DNA sequences in most genomes is protein-coding genes. The vast stretches of the noncoding regions are the regions that regulate the gene activity. These sequences interact with regulatory elements such as transcription factor, chromatin regulators, and noncoding RNAs, which together constitute the epigenome. Epigenomics is the systematic analysis of the global state of gene expression not attributable to mutational changes in the underlying DNA genome. An organism has multiple, cell type-specific, epigenomes comprising epigenetic marks such as DNA methylation, histone modification, and specifically positioned nucleosomes (Stricker, Koflerle, & Beck, 2017). Epigenomic profiling is providing a descriptive view of the chromatin landscape, and data integration enables us to infer functionality from complex datasets. Various sequencing, microarray, and antibody based methodologies are employed to examine the different aspects of epigenetic regulation, including DNA methylation, chromatin accessibility, and histone modifications. Epigenetic analysis techniques called as typing involves profiling of the epigenome. The end-point measurement reflects a proportion or ratio of chromatin with epigenetic marks compared to the total chromatin.

Given the importance of human epigenome, the first project to study the structural and modification of chromatin led to development of the catalog of Encyclopedia of DNA Elements, abbreviated as ENCODE (Davis et al., 2018) and the International Human Epigenome Consortium (IHEC) (Bujold et al., 2016). Public plant epigenomic datasets are emerging quickly, including DNase-seq, ATAC-seq, meDIP-seq, ChIP-seq, and MNase-seq data.

2.4 Beyond genomics and transcriptomics toward proteomics and metabolomics

2.4.1 Proteomics

Proteomics is the study of quantifying proteins in high-throughput manner. In the early 1990s, protein sequencing was done by Edman degradation process. Currently, this is done using both shotgun and targeted approach. Improvements in mass spectrometry (MS) technology have increased sensitivity requiring low concentrations of samples. The high-throughput analyses allow for looking for minimal differences in protein abundances and identifying the posttranslational modifications (Aebersold & Mann, 2016).

Proteomic studies can be done by either chemical labeling or unlabeled techniques. The six major steps included are sample collection, protein extraction, enzymatic digestion of proteins into peptides, separation/fractionation using liquid chromatography (LC) method, peptide and protein identification and quantification by MS, and pathway and network analyses using bioinformatics tools. The field has moved forward from 2D-PAGE-based (dye/fluorescence labeling) protein spot extraction followed by LC-MS or matrix-assisted laser desorption/ionization time-of-flight Ms characterization to more system-wide screening approaches with quantitative steps such as Isotope-Coded Affinity Tagging, Stable Isotope Labeling with Amino Acids in Cell Culture (SILAC), ¹⁸O Stable Isotope Labeling, Isobaric Tagging for Relative and Absolute Quantitation (iTRAQ), and Tandem Mass Tags (TMT) (Bakalarski & Kirkpatrick, 2016) or are label-free (Anand, Samuel, Ang, Keerthikumar, & Mathivanan, 2017; Bantscheff, Lemeer, Savitski, & Kuster, 2012) methods.

Both label-free (Proffitt et al., 2017) and label-based methods such as TMT proteomics from diverse biological matrices have yielded favorable results. The community has not yet built a consensus in terms of data formatting, cleaning, and normalization, for example, the use of ion intensity versus peptide-to-spectrum matches, despite the ongoing efforts through the Proteomics Standards Initiative (Deutsch et al., 2017). Nonetheless, proteomics is advancing our understanding in biomedical research, including diagnosis, protein-based biomarker development, and therapeutics.

2.4.2 Metabolomics

The omics study of metabolites, which are usually the products of biochemical pathways, is called metabolomics. They provide a link connecting genome, transcriptome, proteome to a phenotype. Metabolomics captures small molecule in solid (with solid-state nuclear magnetic resonance (NMR)), liquid (LC-MS), capillary electrophoresis MS (CE-MS), and gas phase MS (GC-MS, tandem-MS). Metabolic analyses includes sample collection, quenching of metabolism, metabolite extraction, chemical derivatization MS, data alignment, filtering, imputation, statistical analysis, annotation, and pathway/network analysis. Depending on the platform used, the sample analysis, data structure, imputation, and normalization scaling differ in data type and instrument. The steps also differ when choosing targeted or untargeted analyses.

2.5 Integrating omics datasets

To understand the actual biological processes in any systems, it is important to integrate all the omics datasets together so that a holistic picture would help the biologists to understand the complex biological pathway(s). The multiomics integration studies have been usually done by conceptual, statistical, and model-based methods. The conceptual methods have provided insights but could result in risky arbitrary connections. Statistical methods from different datasets provide an unbiased integration (Cavill, Jennen, Kleinjans, & Briedé, 2016; Rai, Saito, & Yamazaki, 2017). The model-based integration allows construction of biological pathways or regulatory pathways which could be qualitative or quantitative and these models are considered for hypothesis testing (Rai et al., 2017; Thiele & Palsson, 2010).

Unbiased and element-based integration often uses statistical tools such as clustering, correlation, and multivariate analysis. Clustering approaches such as hierarchical cluster analysis or nonhierarchical methods such k-means clustering are used to identify underlying associations and patterns in the dataset. These methods often produce distinct groupings and provide biological insight. Often, when these approaches are taken to next level with knowledge-based pathways with coexpression and mapping-based approaches, it can provide novel biological insights.

Machine learning (ML) techniques such as Random Forest are used in multiomics experiments such as to identify the regulatory elements, and in studies for specific phenotypic traits such as tuber flesh color, shape, and starch gelatinization. (Acharjee, Kloosterman, Visser, and Maliepaard, 2016). For more complex omics datasets, multivariate analysis allows greater flexibility in experimental design and metadata analysis (Rai et al., 2017) including trends in datasets, and discovery of variance or covariance associations (Meng, Kuster, Culhane, & Gholami, 2014) and topological networks between transcript/protein/metabolite elements (Weckwerth, 2019). Most common multivariate techniques are principal component analysis, partial least squares, and orthogonal projection to latent structures discriminant analysis (Mamat, Azizan, Baharum, Mohd Noor, & Mohd Aizat, 2018; Mazlan, Aizat, Baharum, Azizan, & Noor, 2018; Reinke et al., 2018).

Pathway mapping is a very popular approach that maps different omics datasets to existing metabolic pathways. Several databases are available to study the pathways and some of these tools are listed in Table 2.2. Integrating multiomics dataset can be done by coexpression analysis, which heavily relies on statistical correlations between different omics datasets and to assess the strength of relationships. The relationships are further transformed into weighted network with tools such as weighted gene coexpression network analysis (WGCNA). The WGCNA package is available in R program. These types of analysis have helped scientists to identify hubs and clusters for a pathway of interest and recognize the key regulatory elements in a pathway. Often, the WGCNA approach is followed with Cytoscape visualization (Jiang, Xing, Wang, Zeng, & Zou, 2019; Savoi et al., 2017). Both pathway analysis and coexpression analysis make meaningful integration and are helpful in identifying relationships between different omics datasets. The issues with these pathways are that they are often static and do not take the experimental parameters and perturbation.

The mathematical approach in omics integration aims to develop well-defined differential equations and modeling for a system-level understanding. These analyses involve four steps, which include identification of systems components, understanding the systems regulation and topology, determining mathematical equations, and finally parameter selection and optimization.

The mathematical integration with differential and genome-scale analyses provides a quantitative approach to evaluate the impact of dose of a gene product or a chemical on a particular pathway (Belouah et al., 2019; Voit, 2017;

TABLE 2.2 Summary of databases, functionalities, and license types for the different omics platforms.

Database	Omics	Domain	Functionality	License types
KEGG	<ul style="list-style-type: none"> • Genomics • Transcriptomics • Proteomics • Metabolomics 	Multiple organisms	Biological pathways for processes, diseases, drugs	Open source and licensed
Plant Metabolic Network (PMN)	<ul style="list-style-type: none"> • Genomics • Transcriptomics • Metabolomics 	Plants	Plants specific database containing pathways, reactions	Open source
KBCommons (Knowledge Base Commons)	<ul style="list-style-type: none"> • Genomics • Transcriptomics • Proteomics • Metabolomics • Phenomics • Epigenomics 	Multiple organisms	Platform supporting storing, sharing, analyzing genomics and integrative omics data	Open source
BioCyc database	<ul style="list-style-type: none"> • Genomics • Transcriptomics • Proteomics • Metabolomics 	Multiple organisms	Computationally predicted metabolic pathways and operons (bacteria and archaea). Data support for gene essentiality, regulatory networks, protein features, and GO annotations	Open source
MetaCyc	Metabolomics	Multiple organisms	Pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes	Open source
COVAIN	<ul style="list-style-type: none"> • Transcriptomics • Proteomics • Metabolomics 	Multiple organisms	Workflow including uploading data, data preprocessing, uni- and multivariate statistical analysis, Granger time-series analysis, pathway mapping, correlation network topology analysis and visualization	Open source

Wang et al., 2018). Genome-scale analysis is a mathematical modeling approach that aims to build a genome-scale model and metabolic pathways at the organismal level. This involves primarily four steps such as draft reconstruction using annotated genome, then pathway refinement using experimental results, further network modeling in mathematical format, and lastly, validation and iteration for model accuracy (Thiele & Palsson, 2010). The mathematical models can accurately predict changes or perturbation with database annotation and experimental evidence. However, when the system gets complex especially when dealing with diverse cellular, tissue types, and organelle compartmentalization, the analysis becomes challenging.

2.6 Challenges

Integrating omics experiments is a challenging task as there are too many variables that can make the task complex and ultimately make the interpretation difficult and in some cases result in false positive. The challenges can be grouped into categories listed in Fig. 2.2.

2.7 Machine learning in omics

Huge amounts of data are produced from omics experiment. The problems with these datasets is that they are too large for traditional theoretical and applied statistical methods. This data also has the issue of important signals in a very small region often dominated and masking with noise. For these reasons, the importance of ML and AI methods are getting very popular in extracting valuable information from omics experiments. ML and related deep learning algorithms can handle the large data obtained from NGS and phenotyping platforms for studies addressing precision medicine, precision breeding in agriculture, complex trait dissection, and gene discovery.

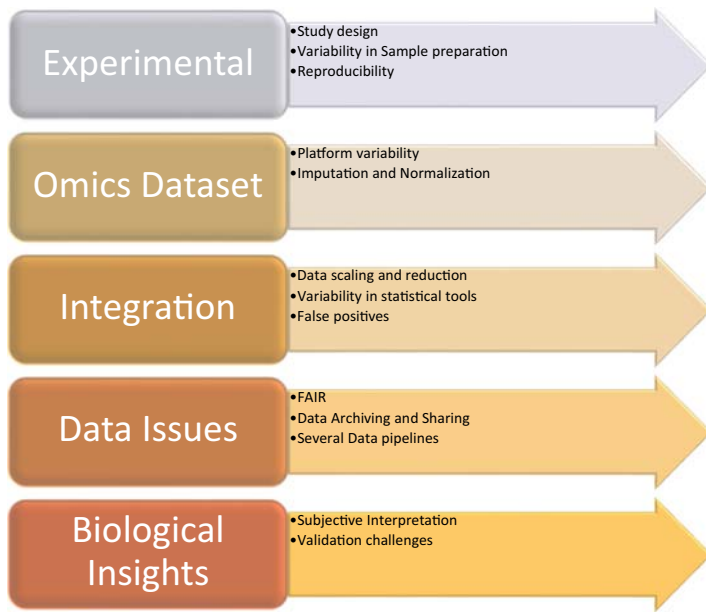


FIGURE 2.2 Data challenges with high throughput Omics experiments.

2.7.1 Machine learning for genomic studies

With the third-generation sequencing technologies, longer reads are produced in comparison to short reads by Illumina sequencing but often are accompanied with the challenge of sequencing errors. To tackle this challenge, Clairvoyante, a deep learning model, was generated using convolution neural network (Luo, Sedlazeck, Lam, & Schatz, 2019). The authors tested Clairvoyante performance to call variants in a genome-wide scenario from 1000 genomes project and found that they achieved 99.67%, 95.78%, and 90.53% F1-score (a measure of test accuracy) when common variants were analyzed, and 98.65%, 92.57%, and 87.26% in whole-genome analysis for Illumina, PacBio, and Oxford Nanopore data. Another popular ML algorithm using artificial neural network is the DeepVariant package. DeepVariant computes the probabilities of three possible allele combinations (homozygous or heterozygous alleles with the reference, and homozygous alleles within the variants) for each variant site, by learning statistical relationships between images of reads around putative variant and true genotype calls. This variant calling method works well for different sequencing technologies (Poplin et al., 2018).

In agriculture, ML tools are getting popular and used in precision agriculture and for smart agriculture. Next-generation phenomics combines precision in trait detection and big data generation by means of high-throughput agri-systems and high-performance computing technologies. The plethora of information from phenomics and genomics data is used in linking and understanding the function of the unknown genes and their network. ML plays a pivotal role for the analysis of complex agricultural data related to plant features and environmental parameters. It allows processing the huge amount of data from sensors and phenotyping platforms, increasing the throughput and accuracy in analysis, as well as its management. The next-generation breeding includes using ML algorithms for precision breeding for the prediction of untested phenotypes in genome selection processes. Random Forest and Bayes models are quite popular ML algorithms used in plant breeding. In the last few years, online sources have been developed for the prediction of genomic estimated breeding values (GEBVs) solGS, a user-friendly online interface implemented in the Nextgen Cassava breeding database (CASSAVABASE, <https://cassavabase.org/solgs>), which allows users to create training populations, input a dataset, and estimate the GEBV of selection candidates. The interactive online exploration and graphical data output makes this tool available to broad number of users (Fig. 2.3).

2.8 Big data storage and management

Handling the deluge of datasets which grows in exponential manner requires high storage and modern and innovative methods. Raw data from sequencing projects are stored in the Sequence Read Archive, which is a repository for short sequence reads (NCBI, <https://www.ncbi.nlm.nih.gov/sra>). Only few agencies, such as European Bioinformatics Institute (EMBL-EBI) and the National Center for Biotechnology Information (NCBI), can store large dataset.

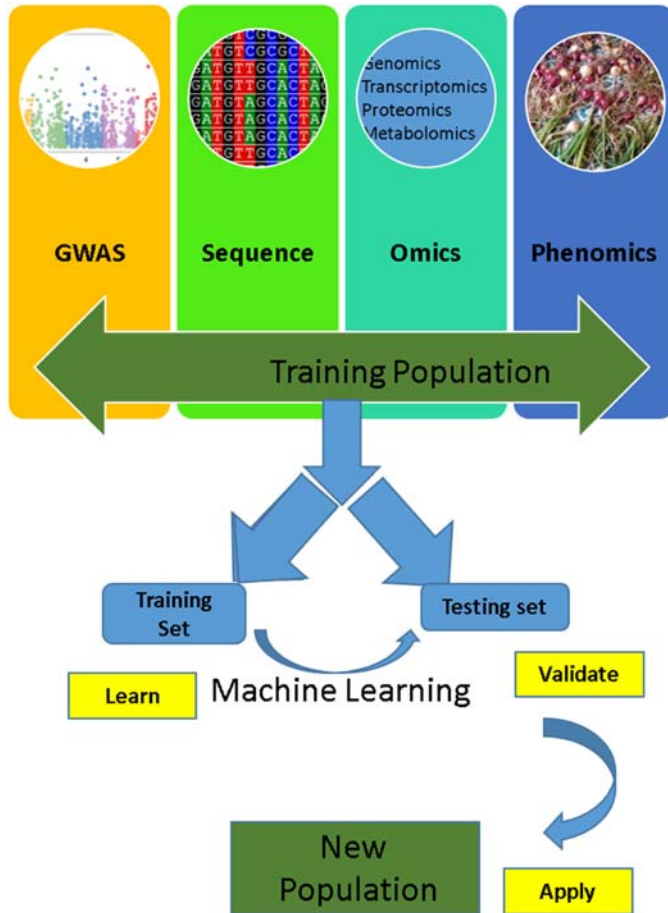


FIGURE 2.3 Machine learning steps with omics datasets as input.

The Amazon S3 storage services are emerging as popular options in terms of cloud-based file system, unlimited capacity, and data security. Beijing Genomic Institute (BGI, Shenzhen) also has built a cloud-based data service for bioinformatics method development, automated analysis, and data delivery.

2.9 Future directions

Predictive models with multiomics datasets hold great potential in biomarker discovery and accelerating drug development process. It is becoming essential to undertake an integrative approach to fully utilize all data types and gain insights into biological systems. ML offers novel techniques to integrate and analyze various omics data, enabling discovery of novel patterns and new biomarkers. With the cloud computing becoming more accessible, there is a future of ML and AI for small- and medium-size institutions and industries.

References

- Acharjee, A., Kloosterman, B., Visser, R. G. F., & Maliapaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinform*, *17*(5), 363–373. Available from <https://doi.org/10.1186/s12859-016-1043-4>.
- Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*, 347–355.
- Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*.
- Anand, S., Samuel, M., Ang, C. S., Keerthikumar, S., & Mathivanan, S. (2017). Label-based and label-free strategies for protein quantitation. In S. Keerthikumar, & S. Mathivanan (Eds.), *Proteome bioinformatics. methods in molecular biology* (vol. 1549). New York, NY: Humana Press.
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*, 166–169.

- Anderson, S. N., Stitzer, M. C., Brohammer, A. B., Zhou, P., Noshay, J. M., O'Connor, C. H., et al. (2019). Transposable elements contribute to dynamic genome content in maize. *The Plant Journal: for Cell and Molecular Biology*, *100*, 1052–1065.
- Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*, 796–815.
- Bakalarski, C. E., & Kirkpatrick, D. S. (2016). A biologist's field guide to multiplexed quantitative proteomics. *Molecular and Cellular Proteomics*, *15*, 1489–1497.
- Bantscheff, M., Lemeer, S., Savitski, M. M., & Kuster, B. (2012). Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, *404*, 939–965.
- Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, *6*, 11, 2015.
- Belouah, I., Nazaret, C., Pétriacq, P., Prigent, S., Bénard, C., Mengin, V., et al. (2019). Modeling protein destiny in developing fruit. *Plant Physiology*, *180*(3), 1709–1724. Available from <https://doi.org/10.1104/pp.19.00086.00086.02019>.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30* (15), 2114–2120.
- Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525–527.
- Bujold, D., Morais, D. A. L., Gauthier, C., Cote, C., Caron, M., Kwan, T., et al. (2016). The international human epigenome consortium data portal. *Cell Systems*, *3*, 496–499.
- Camarena, L., Bruno, V., Euskirchen, G., Poggio, S., & Snyder, M. (2010). Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. *PLoS Pathogens*, *6*.
- Cavill, R., Jennen, D., Kleinjans, J., & Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, *17*(5), 891–901. Available from <https://doi.org/10.1093/bib/bbv090>.
- Chen, J., Hu, Q., Zhang, Y., Lu, C., & Kuang, H. (2014). P-MITE: A database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Research*, *42*, D1176–D1181.
- Coletta, R. F., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biology*, *22*.
- Collard, B. C. Y., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*, 557–572.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13.
- Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.)*, *322*, 1845–1848.
- Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, *46*, D794–D801.
- Deutsch, E. W., Orchard, S., Binz, P. A., Bittremieux, W., Eisenacher, M., Hermjakob, H., Kawano, S., Lam, H., Mayer, G., Menschaert, G., et al. (2017). Proteomics standards initiative: fifteen years of progress and future work. *Journal of Proteome Research*, *16*, 4288–4298.
- Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*, 20140331.
- Fang, Z., & Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics*, *12*, 280–287.
- FastQC. (2017). A quality control tool for high throughput sequence data. [Internet]. Babraham Institute [cited 27.04.17].
- Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *28*, 3169–3177.
- Gage, J. L., Vaillancourt, B., Hamilton, J. P., Manrique-Carpintero, N. C., Gustafson, T. J., Barry, K., et al. (2019). Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. *Plant Genome*, *12*.
- Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*, 875–879.
- Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., et al. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*.
- Grytten, I., Rand, K. D., Nederbragt, A. J., & Sandve, G. K. (2020). Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC Genomics*, *21*, 282.
- Haberer, G., Kamal, N., Bauer, E., Gundlach, H., Fischer, I., Seidel, M. A., et al. (2020). European maize genomes highlight intraspecies variation in repeat and gene content. *Nature Genetics*.
- Hirsch, C. N., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., et al. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *The Plant Cell*, *28*, 2700–2714.
- Hu, Z., Wei, C., & Li, Z. (2020). Computational strategies for eukaryotic pangenome analyses. In H. Tettelin, & D. Medini (Eds.), *The pangenome: diversity, dynamics and evolution of genomes* (p. 2020). Cham: Springer.
- Huang, C., Sun, H., Xu, D., Chen, Q., Liang, Y., Wang, X., et al. (2018). ZmCCT9 enhances maize adaptation to higher latitudes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, E334–E341.
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, *12*, 115–121.
- International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*, *436*, 793–800.

- International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators., Appels, R., Eversole, K., Feuillet, C., Keller, B., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (New York, N.Y.)*, 61.
- Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., & Wessler, S. R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, 431, 569–573.
- Jiang, J., Xing, F., Wang, C., Zeng, X., & Zou, Q. (2019). Investigation and development of maize fused network analysis with multi-omics. *Plant Physiology and Biochemistry: PPB / Societe Francaise de Physiologie Vegetale*, 141, 380–387. Available from <https://doi.org/10.1016/j.plaphy.2019.06.016>.
- Kumar, J., Gupta, D. S., Gupta, S., Dubey, S., Gupta, P., & Kumar, S. (2017). Quantitative trait loci from identification to exploitation for crop improvement. *Plant Cell Reports*, 36, 1187–1213.
- Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, 15, 553.
- Li, C., Song, W., Luo, Y., Gao, S., Zhang, R., Shi, Z., et al. (2019). The HuangZaoSi maize genome provides insights into genomic variation and improvement history of maize. *Molecular Plant*, 12, 402–409.
- Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., Zhang, Z., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, 32, 1045–1052.
- Liu, Y., He, Z., Appels, R., & Xia, X. (2012). Functional markers in wheat: Current status and future prospects. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 125, 1–10.
- Lo, C. C., & Chain, P. S. (2014). Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics*, 15, 366.
- Luo, R., Sedlazeck, F. J., Lam, T. W., & Schatz, M. C. (2019). A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nature Communications*, 10, 998. Available from <https://doi.org/10.1038/s41467-019-09025>.
- Mamat, S. F., Azizan, K. A., Baharum, S. N., Mohd Noor, N., & Mohd Aizat, W. (2018). Metabolomics analysis of mangosteen (*Garcinia mangostana* Linn.) fruit pericarp using different extraction methods and GC-MS. *Plant Omics*, 11(2), 89. Available from <https://doi.org/10.21475/poj.11.02.18.pne1191>.
- Mazlan, O., Aizat, W. M., Baharum, S. N., Azizan, K. A., & Noor, N. M. (2018). Metabolomics analysis of developing *Garcinia mangostana* seed reveals modulated levels of sugars, organic acids and phenylpropanoid compounds. *Scientia Horticulturae*, 233, 323–330. Available from <https://doi.org/10.1016/j.scienta.2018.01.061>.
- Meng, C., Kuster, B., Culhane, A. C., & Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform*, 15(1), 162. Available from <https://doi.org/10.1186/1471-2105-15-162>.
- Nelson, M. G., Linheiro, R. S., & Bergman, C. M. (2017). McClintock: An integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3 (Bethesda)*, 2763–2778.
- Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457, 551–556.
- Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018). Creating a universal SNP and small indel variant caller with deep neural networks. *BioRxiv*. Available from <https://doi.org/10.1101/092890>.
- Proffitt, J. M., Glenn, J., Cesnik, A. J., Jadhav, A., Shortreed, M. R., Smith, L. M., . . . Olivier, M. (2017). Proteomics in non-human primates: Utilizing RNA-Seq data to improve protein identification by mass spectrometry in vervet monkeys. *BMC Genomics*, 18, 877.
- Rai, A., Saito, K., & Yamazaki, M. (2017). Integrated omics analysis of specialized metabolism in medicinal plants. *The Plant Journal: for Cell and Molecular Biology*, 90(4), 764–787. Available from <https://doi.org/10.1111/tpj.13485>.
- Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., et al. (2019). Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, 51, 354–362.
- Reinke, S. N., Galindo-Prieto, B., Skotare, T., Broadhurst, D. I., Singhanian, A., Horowitz, D., et al. (2018). OnPLS-based multi-block data integration: A multivariate approach to interrogating biological interactions in asthma. *Analytical Chemistry*, 90(22), 13400–13408. Available from <https://doi.org/10.1021/acs.analchem.8b03205>.
- Savoi, S., Wong, D. C., Degu, A., Herrera, J. C., Bucchetti, B., Peterlunger, E., et al. (2017). Multi-omics and integrated network analyses reveal new insights into the systems relationships between metabolites, structural genes, and transcriptional regulators in developing grape berries (*Vitis vinifera* L.) exposed to water deficit. *Frontiers in Plant Science*, 8, 1124. Available from <https://doi.org/10.3389/fpls.2017.01124>.
- Schatz, M. C., Maron, L. G., Stein, J. C., Hernandez Wences, A., Gurtowski, J., Biggers, E., et al. (2014). Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biology*, 15, 506.
- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463, 178–183.
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science (New York, N.Y.)*, 326, 1112–1115.
- Smith-Unna, R., Bournnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26, 1134–1144.
- Song, B., Wang, H., Wu, Y., Rees, E., Gates, D. J., Burch, M., et al. (2020). Constrained non-coding sequence provides insights into regulatory elements and loss of gene expression in maize. *BioRxiv*, p. 2020.07.11.192575.
- Springer, N. M., Anderson, S. N., Andorf, C. M., Ahern, K. R., Bai, F., Barad, O., et al. (2018). The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics*, 50, 1282–1288.

- Stricker, S. H., Koflerle, A., & Beck, S. (2017). From profiles to function in epigenomics. *Nature Reviews. Genetics*, 18, 51–66, 10.
- Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1), 93. Available from <https://doi.org/10.1038/nprot.2009.203>.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28, 511–515.
- Van Verk, M. C., Hickman, R., Pieterse, C. M., & Van Wees, S. C. (2013). RNA-Seq: Revelation of the messengers. *Trends in Plant Science*, 18, 175–179.
- Voit, E. O. (2017). The best models of metabolism. *WIREs Systems Biology and Medicine*, e1391
- Wang, J. P., Matthews, M. L., Williams, C. M., Shi, R., Yang, C., Tunlaya-Anukit, S., et al. (2018). Improving wood properties for wood utilization through multi-omics integration in lignin biosynthesis. *Nature Communications*, 9(1), 1579. Available from <https://doi.org/10.1038/s41467-018-03863-z>.
- Weckwerth, W. (2019). Toward a unification of system-theoretical principles in biology and ecology—The stochastic lyapunov matrix equation and its inverse application. *Frontiers in Applied Mathematics and Statistics*, 5, 29. Available from <https://doi.org/10.3389/fams.2019.00029>.
- Yokosho, K., Yamaji, N., Fujii-Kashino, M., & Ma, J. F. (2016). Retrotransposon-mediated aluminum tolerance through enhanced expression of the citrate transporter OsFRDL4. *Plant Physiology*, 172, 2327–2336.
- Zhou, P., Hirsch, C. N., Briggs, S. P., & Springer, N. M. (2019). Dynamic patterns of gene expression additivity and regulatory variation throughout maize development. *Molecular Plant*, 12, 410–425.

Bioinformatics intervention in functional genomics: current status and future perspective—an overview

Swati Sharma¹, Ashwani Kumar¹, Dinesh Yadav² and Manoj Kumar Yadav¹

¹Department of Agricultural Biotechnology, College of Agriculture, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India, ²Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India

3.1 Introduction

The information resulted from postgenomic and high-throughput techniques are no longer a bottleneck in understanding and tackling the biological processes. The biological problems are easy to unravel by sequencing of DNA, proteins using various computational tools, and informatics algorithms for assessing molecular data (Khan, 2018). Bioinformatics is playing a major role in the field of molecular biology ranging from cancer studies in humans to study of microbial pathogens (Katara, 2014). Moreover, to understand the high-throughput techniques such as DNA microarrays, chip-on-chip, protein chips, and recently, the new-generation sequencers, from global prospective, the researchers are handling a vast amount of data generated through these techniques. This huge amount of data generated needs to be analyzed using bioinformatics tools. The first genomic initiative has been set up about 35 years ago, the Human Genome Project, and completed in 2003. Bioinformatics aids in deciphering various human genes and provided information about their structure and organization. A researcher could be able to learn more and more regarding functions of genes and proteins among the similar and dissimilar organisms. The only challenging goal was determining the unit by unit order of nucleotides together making up the human genome (Collins & Fink, 1995). *Arabidopsis thaliana* was the first among the plants and third among the multicellular organism after *Caenorhabditis elegans* and *Drosophila melanogaster*, to be completely sequenced (Tabata et al., 2000). It became the sound basis for further investigations as on completing the sequencing of this plant; it was found that high-throughput technologies will dramatically increase the knowledge on complex biological networks (Hidalgo, 2003). Bioinformatics is an interdisciplinary subject which is the amalgamation of biological and information science that develops new methods and software tools to understand the biological data. It plays a key role to do comprehensive analysis and to understand gene functions with variable levels of protein expression. It is also used to compare the genetic and genomic data and aids to understand various evolutionary aspects of molecular biology. There are various sequence search engines, namely, for homology-based search, NCBI BLAST N and BLAST p.; for orthologous sequence search, Ortho MCL; and for paralogous sequence search, Mc Scan and Mc Scan X are available. Biological databases are used to store and distribute the sequence data, namely, European Molecular Biology Laboratory (EMBL) and the DNA database of Japan (DDBJ). In order to speed up the analysis, bioinformatics enriched itself with a lot of resources, facilities, and databases which are updated timely with new information and knowledge. This review enlightens various bioinformatics methods to solve the biological problems which are related to functional genomics.

3.2 Functional genomic approaches

Functional genomics may be referred to as the development and application of global (genome-wide or system-wide) experimental and systematic approaches that help to assess the gene function by use of information provided by structural genomics (Bouchez & Höfte, 1998). It deals with the study of genes and intergenic regions of the genome which

contributes to the different biological processes. The main goal of functional genomics is to generate a particular phenotype with the help of different components of a biological system. Some functional genomic approaches are mainly based on DNA level (genomics and epigenomics), RNA level (transcriptomics), protein level (proteomics), and metabolite level (metabolomics).

3.3 Serial analysis of gene expression

Serial analysis of gene expression (SAGE) is a unique method and used for identification of transcripts and quantification of eukaryotic genome. The basic principle for this is the determination of a normal gene structure and identification of structural changes in an abnormal genome (Wang, 2004). It is mainly based on representing the mRNAs by using a short sequence tags followed by the concatenation of tags for cloning to allow the sequencing analysis. This technique does not require prior knowledge of gene of interest. Velculescu, Zhang, Vogelstein, and Kinzler (1995) developed a high-throughput method of determining the absolute effluence of every transcript in population of cells (Fig. 3.1). mRNA obtained from cells allows to convert in double-stranded DNA form. Digestion was performed with a 4-bp cutter “anchoring enzyme” NlaIII and then the poly-A proximal ends collected and ligated to a linker fragment. The mentioned linker fragment harbors a 5'-GGGAC-3' sequence, which is the site of recognition of the Type IIS restriction endonuclease BsmFI. It cleaves the cDNA 15 bp away in the 3' direction from the recognition site. A 15-bp long

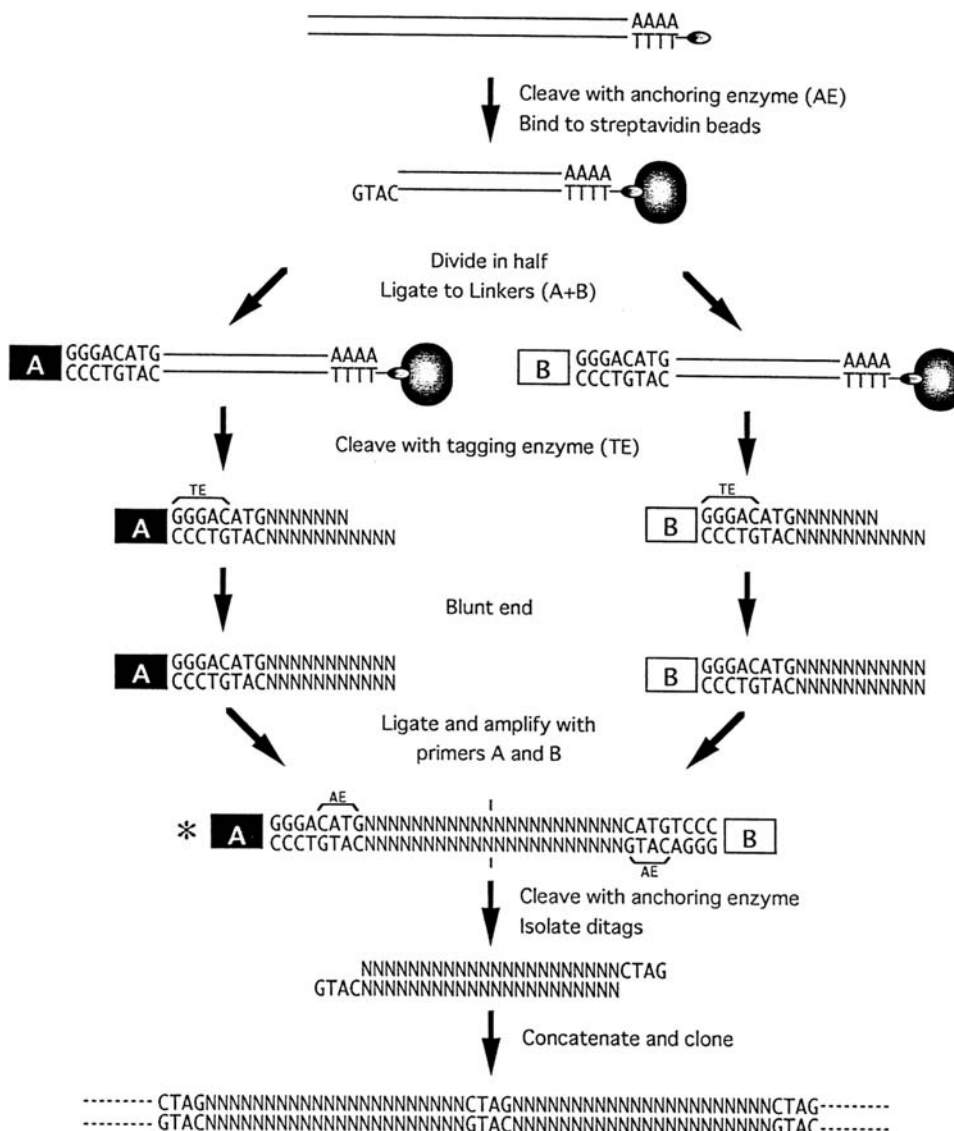


FIGURE 3.1 A SAGE procedure. The AE used is *Nla*III and TE used in the procedure is *Bsm*FI. Boxes A and B are the independent linkers, 39 portions of which are designed to contain TE sequence. Transcript-derived tag sequences are denoted by Ns. Blunt end ligation step is denoted as *, and discussed later in the text. AE, anchoring enzyme; SAGE, Serial analysis of gene expression; TE, tagging enzyme. Adapted from Yamamoto, M., Wakatsuki, T., Hada, A., & Ryo, A. (2001). Use of serial analysis of gene expression (SAGE) technology. *Journal of Immunological Methods*, 250(1–2), 45–66.

fragment called tag released by treatment of the linker-ligated cDNA with BsmFI from a defined position of each cDNA. The tags are concatenated and cloned into a plasmid vector, which is then sequenced after removal of this linker fragment. Generally, for a given sample, around 10,000–100,000 tags may be analyzed. The profusion of the transcript which corresponds to the tag is represented by the number of each tag in the total sample. The next main step is to identify the gene which corresponds to the tag or tag annotation. The 15-bp tag sequence is generally used as a query to search expressed sequence tags (ESTs) or cDNA databases of any organism of interest through BLAST search (Altschul, Gish, Miller, Myers, & Lipman, 1990). Results of tag counts and tag annotation are then combined finally into a gene expression profile. Gene expression profiles are then compared of two samples that are treated differently, then we will be able to tell which gene is up- or downregulated in response to the particular treatment. In short, following are the steps to the SAGE procedure:

- mRNA of an input sample (e.g., a tumor) isolated.
- Remove a small portion of sequence of mRNA molecule which is used for analysis.
- Link these small sequences together to form a longer chain or concatamer.
- Clone these chains into a vector which can be taken up by bacteria.
- Then sequence the chains with the help of high-throughput sequencer.
- Processing of data to count the small sequence tags with the help of a computer.

USAGE, a web-based application which comprises a set of tools to compare and analyze SAGE data. USAGE is accessible at <http://www.cmbi.kun.nl/usage> free of cost for academic institutions. In addition, it enhances the functionality and flexibility of data (Van Kampen et al., 2000). Some of the SAGE databases are:

1. SAGE net: This is the database known as SAGENet (<http://www.sagenet.org>) which is maintained by the Vogelstein/Kinzler Lab at Johns Hopkins. It is used mainly for colon cancer, pancreatic cancer, and some normal tissues of these cells.
2. SAGEmap: This is developed by National Institute of Health's (NIH) National Centre for Biotechnology Information (NCBI) and NIH's Cancer Genome Anatomy Project (CGAP). This database is considered as a public gene expression repository and unique in many ways.
3. Genzyme's SAGE database: Database is used to create SAGE tag libraries for contracting parties. This database is also available through other agencies such as Celera Genomics and Compugen.

Besides this, few other SAGE analysis tools are available such as SAGE300. The SAGE data is obtained with the help of sequencing the short DNA tags, although data may have errors due to sequencing (Tuteja & Tuteja, 2004).

3.3.1 Advantages of serial analysis of gene expression

1. SAGE studies may be proved to be an effective tool in human cancer studies with the help of the gene expression profile studies from cancer and normal tissue of interest. A large number of genes recognized as tumor-specific genes. Northern blot analysis has been done to confirm the differential expression of related gene (Yamamoto et al., 2001).
2. SAGE technique is very much helpful in the areas such as cardiovascular biology, stem cell biology, cardiovascular development, angiogenesis, atherosclerosis, and lipid regulation. It is mainly due to the electronic nature of SAGE databases. Direct comparison of libraries may be done by different investigators. CGAP genome annotation initiative may be used for gene expression queries regarding human heart SAGE library (Patino et al., 2002).
3. SAGE analysis may be done in immunological studies for human monocytes, macrophages, and their differentiated descendants. By comparing the SAGE profiles of related cells, it was discovered that granulocyte macrophage-colony stimulating factor (GM-CSF)-induced and M-CSF responsible macrophages expereed comparable sets of genes and expressed similar sets of genes, implying functional similarity (Chen, Centola, Altschul, & Metzger, 1998).

3.3.2 Drawbacks of serial analysis of gene expression technique

1. It does not compute the authenticity of expression level of a gene.
2. The size of a tag obtained after SAGE analysis is 10 bases, making it difficult to assign a tag to a specific transcript with accuracy.
3. Two different genes could have the same tag and the same gene that is alternatively spliced could have different tags at the 3' ends.

- The mRNA transcript allocated with each tag could be made even more arduous and uncertain on interpolating the sequencing errors into the process.

3.4 DNA microarray

DNA microarrays comprise various microscopic DNA spots (probes) confined to a solid surface, namely, glass or a silicon chip or microscopic beads (Illumina). Under high stringency conditions, from any sample of interest, single-stranded DNA that is labeled or antisense RNA fragments are hybridized to the DNA microarray. DNA microarray pinpoints the probe using its location revealing the amount of hybridization detected which is equivalent to the level of nucleic acids from the commensurating location among the original sample in genome (Bunnik & Roch, 2013) (Fig. 3.2).

3.4.1 Applications of microarray

- Microarray aids in examining the huge amount of former or current samples. Also, it has been proved to be efficacious in estimating the role of a certain marker in tumors.
- DNA microarray analyzes the whole bacteria genome viability using a small amount of DNA as there is an immense increase in resistant bacteria leading to casual infections causing failure of antibiotics (Govindarajan et al., 2012).
- Drug target characterization, identification, and selection.
- Cellular response to bacterial infection.
- It diagnoses the presumed genetic disease by testing the existence of mutations.

3.4.2 Drawbacks of microarray

- DNA microarray traces various samples simultaneously but it is a complicated procedure.
- Despite of being a popular technology working for more than thousands of genes, it requires proficiency and skills for data normalization and analysis.
- Also, the technique works for only predefined sequences.
- The technique is based on hybridization but it necessitates the high-power computing facilities.

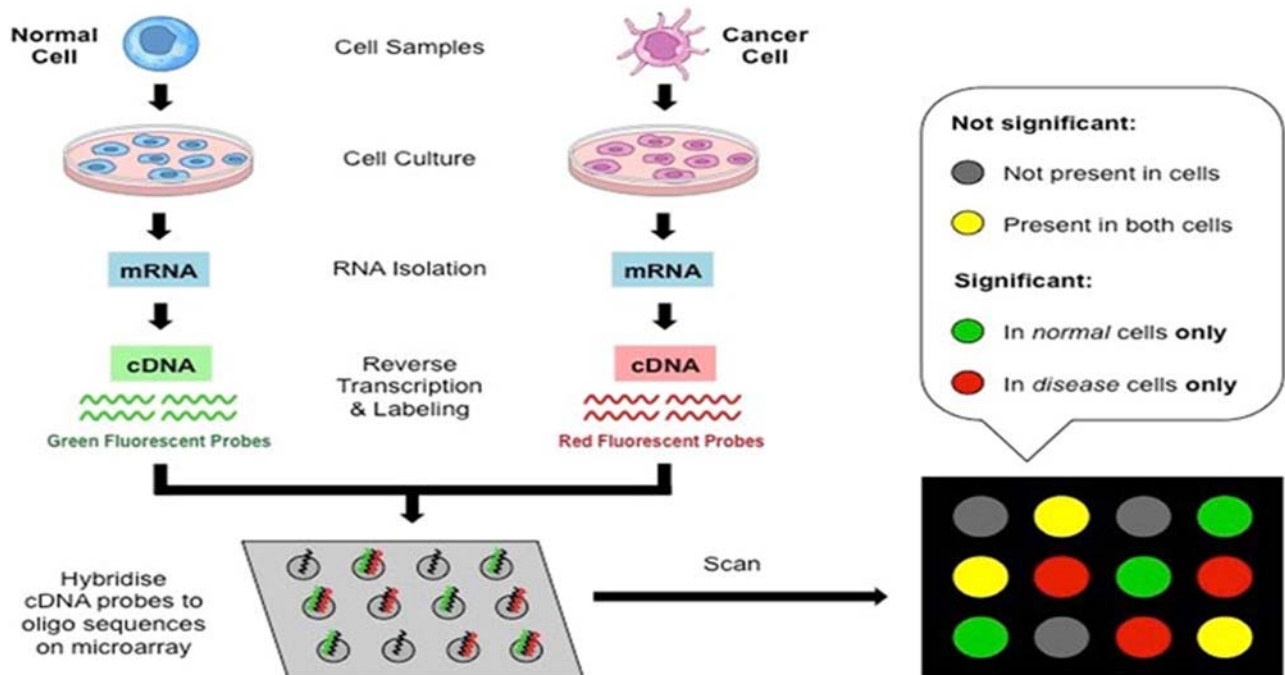


FIGURE 3.2 Schematic representation of steps of microarray.

3.4.3 Bioinformatics tools for microarray data analysis

The data collected by the microarray experiment generates extremely huge files, which are examined for the results. In order to make the process easier, a variety of software has been developed. The Affymetrix GeneChip platform, for example, is one of the most widely used software for studying gene expression. Following are the software for Affymetrix data analysis.

3.4.3.1 GeneChip Operating Software

It works in hardware management, image analysis, expression assessment, and data normalization. Also, it performs the normalization and estimates quality control parameters.

3.4.3.2 Affymetrix Expression Console Software

The software summarizes the probe sets with enumerating and normalizing the expression arrays of gene chips. The software is rigged with Microarray Suite 5.0 (MAS5) normalization algorithm, Probe Logarithmic Intensity Error Estimation normalization, and Robust Multichip Analysis (RMA) normalization.

Moreover, following are some of the free software for academic use:

RMA Express: Robust Multichip Average, a program used to assess the gene expression summary values for Affymetrix GeneChip. The software is free for academic use and can be downloaded from <http://rmaexpress.bmbolstad.com>. RMA normalization can also be performed using R (<http://www.r-project.org>) and Bioconductor (<http://www.bioconductor.org>).

dCHIP: Initially, Cheng Li and Wing Hung Wong evolved the DNA-Chip Analyzer (dCHIP) by executing a model-based expression analysis for Affymetrix gene expression arrays. The Affymetrix raw data (dat and cel files) and processed data (quantified expression values as a tab delimited file) could easily be processed by this software. A large data analysis such as SNP array, exon array, and tiling arrays can also be done.

The other features of the software are normalization and quality control, hierarchical clustering and comparison of samples.

Few other software which are easy and free to access are SNOMAD (web-based tool), TM4 (Spotfinder, Microarray Data Analysis System), Genesis, Gene Expression Model Selector, etc. (Mehta & Rani., 2011).

MIAME (minimum information about a microarray experiment): In order to report the microarray experiments, FGED society fabricated this standard that specifies the required information for elucidating the experiment results evidently. More precisely, it illustrates the required information to certify the interpretation of microarray data at ease leading to the development of data analysis tools. Various public databases such as ArrayExpress and Stanford Microarray Database are storing gene expression data using the MIAME standard, including the Gene Expression Omnibus. These databases in this age dispense some additional facilities for data analysis and annotation purposes (Brazma et al., 2001; Kremer et al., 2001).

3.5 Next-generation sequencing technologies

The three most prominent and foremost next-generation sequencing (NGS) platforms, namely, Roche 454 platform (Roche Life Sciences), the Applied Biosystems SOLiD platform (Applied Biosystems), and Illumina (previously known as Solexa) Genome Analyzer, and HiSeq platforms (Illumina), are used at large scale.

3.5.1 Illumina sequencing

Bruno Canard and Simon Sarfati at the Pasteur Institute in Paris innovated this technique at first. Although, Shankar Balasubramanian and David Klenerman of Cambridge University established this and consequently founded *Solexa*, a company later acquired by *Illumina*. The method is based on the ability of single-dye terminators to identify the single bases when introduced into DNA strands. Reversible termination sequencing technology is a *sequencing-by-synthesis approach* that concludes the template sequence by stepwise primer elongation. On Illumina platform, it is generalized as a second-generation sequencing technology.

Ion Torrent sequencing is based on the detection of hydrogen ions that are released during the polymerization of DNA and sequence DNA based on a semiconductor chip that is released in February 2010. Also, it is named as Ion Torrent sequencing, pH-mediated sequencing, silicon sequencing, or semiconductor sequencing.

3.5.1.1 Cost of sequencing full genome

1. In June 2009 *Illumina* announced Personal Full Genome Sequencing Service at \$48,000 per genome.
2. In November 2009 *Complete Genomics* sequences a complete human genome for \$1700.
3. In May 2011 *Illumina* lowered its Full Genome Sequencing service to \$5000 per human genome, or \$4000 if ordering 50 or more.
4. Several companies, namely, *Life Technologies* in January 2012, *Oxford Nanopore Technologies* in February 2012, and *Illumina* in February 2014, started to claim that as the cost of sequencing begins to decline, their equipment will achieve \$1000.

3.5.2 Applications of next-generation sequencing

1. The exact order of nucleotide occurrence in DNA could be attained by sequencing methods. The genetic information can be elucidated from any biological system using DNA sequence. F. Sanger in 1975 developed the Sanger sequencing method which was the first generation method of sequencing to be developed. There were certain limitations to the method inherent in nature regarding throughput, speed, scalability and its resolution, second-generation of sequencing method, or NGS developed in order to fulfill the uprising demand of a sequencing method which is cheaper as well as faster in technology.
2. Principally, the basic idea behind NGS is based on the sequencing of thousands of fragment of DNA using a single sample, also known as massive parallel sequencing. It allows the large stretch of DNA base pairs to be sequenced which in results produces hundreds of gigabases of data in single sequential run.
3. The third-generation sequencing method has been developed but it is not as mature as the second-generation sequencing method (Hayden, 2009), therefore being infant, it could not be widely accepted till now, but the NGS methods really are.
4. Molecular biology: NGS plays a vital role in molecular biology while studying the whole genome and encoded proteins. The information retrieved regarding changes in genes and their alliance and affiliation with various diseases and phenotypes helps researchers to learn. Also, it helps in identification of drug targets.
5. *Evolutionary biology* aids in estimating the correlation between the organisms and their development.
6. *Medicine*: The presence of any genetic disease-related risk could be decided, if any, using sequencing methods by the medical technicians.
7. *Forensics*: The use of DNA sequencing methods has been established in DNA profiling and paternity tests in field of identification of forensics. Various samples such as fingerprints of any organism, hairs, saliva, etc. are used as samples in estimating the different separating DNA patterns which is the basis of identification. A certain unique pattern using a single strand could be produced by detecting specific genome as each and every living organism comprises a unique DNA and could be determined via DNA testing. No two individual shares the exact similar DNA pattern, if any, a rare case.

However, NGS methods are much capable as they cope up with the traditional methods (Sanger sequencing) by providing a faster alternative to them. NGS ensures to be very fast as a whole genome in a single day could be sequenced by researchers. For example., *Illumina*, which costs less than \$5000 per genome could sequence more than five human genomes in a single run, resulting into generation of data within a week. The genes including their regulatory pathways associated with diseases could be determined by using high-throughput sequencing (HTS) method.

Exome sequencing reveals the disease-related variations and mutations in exome region. It helps to determine the coding regions of protein within the genome.

Targeted resequencing computes the level of sequencing among the genomic region of interest. Being a small subset such as exome, an advantage using targeted resequencing is that it does not involve higher sequencing cost.

Chromatin immunoprecipitation sequencing (ChIP-Seq): The interaction among protein, DNA and RNA is analyzed using this method. It enables the identification of the binding sites of the DNA associated proteins. Also, it interprets various regulation events such as gene regulation, DNA repair, and DNA synthesis.

RNA sequencing (RNA-seq): It is a transcriptome sequencing approach which comprises functions such as transcript analysis and detection with low expression levels and with or without reference sequence, respectively. Moreover, the method is found to be more precise in quantifying the exact expression levels.

3.5.3 Bioinformatics tools for next-generation sequencing

TopHat: It is an open-source software which helps in the alignment of reads among RNA-seq to the reference genome. It does not rely over the splice sites (Lee et al., 2012).

Bambino: It is a viewer for next-generation sequence files (Edmonson, Zhang, & Yan, 2011).

Tablet: It is Java based and available for Linux, OSX, Windows, and Solaris platforms, in both 32- and 64-bit versions. It provides a sequence level as well as contig overview. Also, it is more capable in highlighting the disagreements among the reference or consensus sequence in the mapped reads.

The Integrative Genomics Viewer: It is an open-source visualization tool (<http://www.broadinstitute.org/igv/>) which aids to explore huge scale of data sets of genome. A variety of array-based data have been supported, namely, expression and copy-number arrays, RNA interference screens, methylation, genomic annotations, and gene expression.

The Savant (Sequence Annotation, Visualization and ANalysis Tool) Genome Browser: It is an open-source desktop visualization and analysis browser developed for visualizing and analyzing genomic data, including the HTS data, for example, NGS, with low memory requirements.

Magic Viewer: It was evolved to align short read visualization and annotation.

Geneious: It is an analysis tool to visualize sequence and a number of operations applied for visualizing and manipulating next-generation sequence data. It also provides tools for the assembly, alignment, and annotation of genomic reads and sequence with exploratory alignment against public repositories using the BLAST sequence search capability.

Mass spectroscopy: Orbitrap is the most forward mass spectrometer available till date with a high resolution, a high mass accuracy, and a large dynamic range, making it convenient to be applied to the proteomic and metabolomic applications.

3.6 Databases and genome annotation

Genome annotation is based on the assessment of functional elements among the genomic sequence. The sequencing of DNA leads to produce the sequences of unknown function (Abril & Castellano Hereza, 2019). Genome annotation results into the determination of the function of the product of a predicted gene via in silico method. For this to happen, several necessary features of bioinformatics software must include (1) signal sensors (e.g., for TATA box, start and stop codon, or poly-A signal detection); (2) content sensors (e.g., for G + C content, codon usage, or dicodon frequency detection); and (3) similarity detection (e.g., between proteins from closely related organisms, mRNA from the same organism, or reference genomes) (de Sá et al., 2018). Biological databases fulfill the requirements.

3.6.1 Biological databases

The biological databases fall under different categories: (1) DNA, (2) RNA, (3) protein, (4) expression, (5) pathway, (6) disease, (7) nomenclature, (8) literature, (9) standard, and (10) ontology (Zou, Ma, Yu, & Zhang, 2015) (Fig. 3.3).

On the basis of source, there are two types of database: **primary and secondary**.

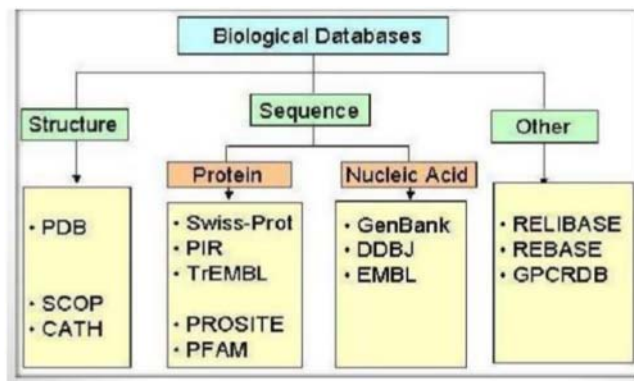


FIGURE 3.3 Types of biological databases. Adapted from NCBI.

3.6.1.1 Primary database

The primary databases contain biomolecular data in its original form. EMBL, GenBank, DDBJ, SWISS-PROT, TREMBL, and PIR constitute the primary databases.

3.6.1.1.1 DNA databases

GenBank is one of the representative of DNA databases as of December 2014, comprising over 184 billion nucleotide bases in 179 million sequences or more. DNA databases establish the reference genome (e.g., NCBI RefSeq), human genetic variation profiling (e.g., dbSNP), and association of genotype with phenotype (e.g., EGA) and help to identify the human microbiome metagenomes (e.g., IMG/HMP) (Zou et al., 2015). The human DNA databases assemble the reference genome (e.g., NCBI RefSeq) and human genetic variation profiling (e.g., dbSNP) and associates the genotype and phenotype together (e.g., EGA) and microbiome metagenomic identification of humans.

EMBL: It was established by collaboration of GenBank and DDBJ.

DDBJ: DNA Data Bank of Japan used to collect DNA sequences.

SWISS-PROT: It is a *protein database* that consists of about 547,357 proteins annotated manually in January 2015 and aids in providing minimum redundancy and higher integration with other databases. protein data bank (PDB) (established in 1971) as determined by X-ray crystallography and nuclear magnetic resonance (NMR) is the other example of protein database for determining 3D structures of biological macromolecules. As of December 30, 2014, PDB comprises 105,465 biological macromolecular structures where 27,393 entries belong to human.

3.6.1.1.2 RNA databases

For decoding ncRNAs, the human RNA databases are constructed (e.g., GENCODE) (Consortium, 1., 2012), specifically lncRNAs attracting the current interest (e.g., LncRNAWiki). RNA central is one of the representative examples of RNA database. It avails the unified access to the ncRNA sequence data supplied by various number of multiple databases such as Rfam, lncRNAdb, and miRBase. (<http://rnacentral.org>) (Table 3.1).

3.6.2 Functional genomic databases

These databases provide information about the functions of genes for example., Databases used for information retrieval system, that is, BLAST, commonly used by the scientist for predicting and analyzing the information regarding function of new or unknown genes. The foremost dedicated genomic databases are described in the following sections.

TABLE 3.1 The biological information and the type of source.

S. no.	Type of information	Source
1.	Nucleotide sequence	GenBank (http://www.ncbi.nlm.nih.gov/genbank/) EMBL (http://www.ebi.ac.uk/embl/) DDBJ (http://www.ddbj.nig.ac.jp)
2.	Nonredundant EST sequence	UniGene (http://www.ncbi.nlm.nih.gov/unigene) TIGR Gene Indices (http://www.tigr.org/tdb/tgi)
3.	Protein sequence and annotation	Uniprot (http://www.uniprot.org/)
4.	Protein structure	PDB, (http://www.rcsb.org/pdb)
5.	Metabolic pathway	KEGG (http://www.genome.ad.jp/kegg/)
6.	Gene expression (cDNA microarray) data	GEO (http://www.ncbi.nlm.nih.gov/geo/) ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) SMD (http://smd.princeton.edu/)
7.	Database of essential genes for prokaryotes and eukaryotes	DEG (http://tubic.tju.edu.cn/deg/)

EST, Expressed sequence tag.

Source: Adapted from Katara, P. (2014). Potential of Bioinformatics as functional genomics tools: an overview. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 3, 52.

3.6.2.1 Rice functional genomics

KOME database (Knowledge-based Oryza Molecular Biological Encyclopedia) gathers about 38 000 full-length cDNAs of *japonica* cv. Nipponbare. A number of 10,081 and 12,727 full-length cDNA sequences from Gaungluai 4 and Minghui 63, respectively, comprised by the rice *indica* cDNA database (RICD) database. *Affymetrix GeneChip Rice Genome Array* examines the expression profiles in various stressfull conditions in elite hybrid rice *Shanyou 63* and its parents *Zhenshan 97* and *Minghui 63*, present in the information platform of Collection of Rice Expression Profiles (CREP). The comparison between transcriptomes of super hybrid rice *LYP9* and its parental cultivars 93–11 and PA64s was performed using gene expression microarrays. *Affymetrix GeneChip Rice Genome Array* determines the quantitative trait loci (eQTLs) expression in rice seedlings and flag leaves during heading period using recombinant inbred lines, which was developed by performing a cross between *Zhenshan 97* and *Minghui 63* (Wang et al., 2010; Wei et al., 2014).

The functional genomics of rice research has enriched the resources with genes such as *Xa21* and *xa13* conferring resistance to plants against rice bacterial leaf blight. *Pigm* and *Bsr-d1* could also be used as a breeding source for disease resistance specifically to rice blast. Wild rice also consisting of a gene *Bph 14* identified originally in *Oryza minuta* for obtaining resistance against brown planthopper. The local varieties also contributed by developing various alleles, such as brown plant hopper resistance gene *BPH3*, salt tolerance gene *HKT2*, submerge tolerance gene *Sub1*, and high-temperature tolerance gene *OstT1*. The genes have a huge potential for breeding in rice.

Some of the databases for the molecular plant are *IC4R* (<http://ic4r.org/>), *RICD* (<http://202.127.18.221/ricd/index.html>), *TIGR* (<http://rice.plantbiology.msu.edu/>), *IRRI* (<http://irri.org/>), *CREP* (<http://crep.ncpgr.cn/>), etc. (Li et al., 2018).

3.6.2.2 Functional genomics in Malvaceae family plants

Several economically flowering plant species constitute the category such as cotton, cacao, and durian. Ma-Gen Db was developed as a user-friendly database for decoding and as functional genomic hub for this plant community, available at <http://magen.whu.edu.cn>. There is an availability of eight types of 367 deep-sequencing data for 13 species. The database aids the generation of multiple dynamic charts and hyperlinks. All the functional annotations for gene, transcript, and protein displayed on a page are named as Genewiki. MaGenDB is a database where a total number of 374 processed omics data of nine techniques with 18 types of annotation and more than 24 million functional elements are stored and conferred in a user-friendly way using well-designed custom dynamic charts. In a concluding note, the database is filling out the gap for a salient plant family and, thereby, generating an functional comparison system (Wang et al., 2020).

3.6.2.3 Functional genomics in fungi

Fungi database (available at <http://FungiDB.org>) is a functional genomic resource which was developed with the partnership with the NIAID-funded Eukaryotic Pathogen Bioinformatics Resource Centre (<http://EuPathDB.org>). The database consisting of the genome sequence and annotation from 18 species from several classes, including Ascomycota, Eurotiomycetes, Sordariomycetes, Saccharomycetes, and Basidiomycota, Pucciniomycetes and Tremellomycetes, and the basal “Zygomycete” lineage Mucormycotina. FungiDB enlightens various functional genomic data sets (1) for *Aspergillus flavus*, *Aspergillus terreus*, *Aspergillus niger*, and *Gibberella moniliformis*. EST is data retrieved from dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>). (2) Based on different synchronization methods, cell cycle microarray data is derived for *Saccharomyces cerevisiae*. (3) RNA-sequence data is derived from *Rhizopus oryzae* during hyphal growth and (4) two hybrid yeast data are obtained from *S. cerevisiae* (Stajich et al., 2012).

Some other databases for the study of genome are *AgBase* database for functional genomic resource, available at (<http://www.agbase.msstate.edu/>); for studying diversity among Rubiaceae family, *MoccaDB* database is available (<http://moccadb.mpl.ird.fr/>); the other one, *TFGD* database is used for tomato functional genomic databases (<http://ted.bti.cornell.edu/>); *SFGD* database is for soybean functional genomic database (<http://bioinformatics.cau.edu.cn/SFGD/>), etc.

3.7 Conclusion

The genomic data resulting from sequencing created various huge challenges as well as several opportunities to study the genomes of organism. The bioinformatic tools mentioned in the present review article including databases and software play an efficient role in handling out those challenges. Several functional genomic approaches with their databases are mentioned to tackle the biological problems generating from the huge size of data. Although the functional genomic databases are continuously updated with mined knowledge and new information in order to provide much more reliable information for genomics-related analysis.

References

- Abril, J. F., & Castellano Hereza, S. (2019). *Genome annotation* (pp. 195–209). Elsevier.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.
- Bouchez, D., & Höfte, H. (1998). Functional genomics in plants. *Plant Physiology*, 118(3), 725–732.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., . . . Roch, K. G. (2013). An introduction to Functional Genomics and System Biology. *Advances in wound care.*, 2(9), 490–498.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C. & Gaasterland, T. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4), 365–371.
- Bunnik, E. M. & Le Roch, K. G. (2013). An introduction to functional genomics and systems biology. *Advances in wound care*, 2(9), 490–498.
- Chen, H., Centola, M., Altschul, S. F., & Metzger, H. (1998). Characterization of gene expression in resting and activated mastcells. *The Journal of Experimental Medicine*, 188, 1657–1668.
- Collins, F. S., & Fink, L. (1995). The Human Genome Project. *Alcohol Health and Research World*, 19(3), 190–195.
- Consortium, I. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56–65.
- de Sá, P. H., Guimarães, L. C., das Graças, D. A., de Oliveira Veras, A. A., Barh, D., Azevedo, V., . . . Ramos, R. T. (2018). *Next-generation sequencing and data analysis: Strategies, tools, pipelines and protocols. Omics Technologies and Bio-Engineering* (pp. 191–207). Academic Press.
- Edmonson, M. N., Zhang, J., Yan, C., et al. (2011). Bambino: A variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics (Oxford, England)*, 27, 865–866.
- Govindarajan, R., Duraiyan, J., Kaliyappan, K., & Palanisamy, M. (2012). Microarray and its applications. *Journal of Pharmacy & Bioallied Sciences*, 4(Suppl 2), S310.
- Hayden, E. C. (2009). Genome sequencing: the third generation. *Nature*, 457(7231), 768–769.
- Hidalgo, O. B. (2003). Functional genomics and bioinformatics: an overview. *Bioteconología Aplicada.*, 20(3), 183.
- Katara, P. (2014). Potential of Bioinformatics as functional genomics tools: An overview. *Network Modeling Analysis in Health Informatics and Bioinformatics.*, 3, 52.
- Khan, N. T. (2018). Structural and Functional Bioinformatics. *Letters in Health and Biological Science*, 3(1), 7–11.
- Kremer, S., Stewart, J., Taylor, R., Vilo, J., & Vingron, M. (2001). Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics*, 29, 365–371.
- Lee, H. C., Lai, K., Lorenc, M. T., Imelfort, M., Duran, C., & Edwards, D. (2012). Bioinformatics tools and databases for analysis of next-generation sequence data. *Briefings in Functional Genomics.*, 11(1), 12–24.
- Li, Y., Xiao, J., Chen, L., Huang, X., Cheng, Z., Han, B., & Wu, C. (2018). Rice functional genomics research: past decade and future. *Molecular plant.*, 11(3), 359–380.
- Mehta, J. P., & Rani, S. (2011). *Software and tools for microarray data analysis in Gene Expression Profiling* (784, pp. 41–53). Humana Press.
- Patino, W. D., Mian, O. Y., & Hwang, P. M. (2002). Serial analysis of gene expression: technical considerations and applications to cardiovascular biology. *Circulation Research*, 91(7), 565–569.
- Stajich, J. E., Harris, T., Brunk, B. P., Brestelli, J., Fischer, S., Harb, O. S., & Stoeckert, C. J., Jr (2012). FungiDB: an integrated functional genomics database for fungi. *Nucleic Acids Research*, 40(1), 675–681.
- Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T. & Kawashima, K. (2000). Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, 408(6814), 823–826.
- Tuteja, R., & Tuteja, N. (2004). Serial analysis of gene expression (SAGE): unraveling the bioinformatics tools. *Bioessays.*, 26(8), 916–922.
- Van Kampen, A. H., van Schaik, B. D., Pauws, E., Michiels, E. M. C., Ruijter, J. M., Caron, H. N., & van Der Mee, M. (2000). USAGE: A web-based approach towards the analysis of SAGE data. *Bioinformatics (Oxford, England)*, 16(10), 899–905.
- Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science (New York, N.Y.)*, 270, 484–487.
- Wang, D., Fan, W., Guo, X., Wu, K., Zhou, S., Chen, Z., . . . Zhou, Y. (2020). MaGenDB: a functional genomics hub for Malvaceae plants. *Nucleic Acids Research.*, 48(1), 1076–1084.
- Wang, L., Xie, W., Chen, Y., Tang, W., Yang, J., Ye, R., Liu, L., Lin, Y., Xu, C., Xiao, J., et al. (2010). A dynamic gene expression atlas covering the entire life cycle of rice. *The Plant Journal: for Cell and Molecular Biology*, 61, 752–766.
- Wang, S. M. (2004). Understanding SAGE data. *Trends in Genetics.*, 23(1), 42–50.
- Wei, L., Gu, L., Song, X., Cui, X., Lu, Z., Zhou, M., Wang, L., Hu, F., Zhai, J., Meyers, B. C. , et al. (2014). Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 3877–3882.
- Yamamoto, M., Wakatsuki, T., Hada, A., & Ryo, A. (2001). Use of serial analysis of gene expression (SAGE) technology. *Journal of Immunological Methods*, 250(1–2), 45–66.
- Zou, D., Ma, L., Yu, J., & Zhang, Z. (2015). Biological databases for human research. *Genomics, proteomics & bioinformatics.*, 13(1), 55–63.

Genome informatics: present status and future prospects in agriculture

Pramod Kumar Yadav¹, Rahul Singh Jasrotia¹ and Akanksha Jaiswar²

¹Department of Computational Biology & Bioinformatics, JIBB Sam Higginbottom University of Agriculture, Technology & Sciences, (Formerly AAI-DU), Prayagraj (Allahabad), Uttar Pradesh, India, ²Centre for Agricultural Bioinformatics (CABin), ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

4.1 Introduction

Genome informatics (or geninformatics) is the new emerging discipline of bioinformatics, where computational and statistical techniques are applied to study the structure and function of genes and genome of an organism (Fadiel et al., 2005). Watson and Crick (1953) proposed the double-helix model of DNA (deoxyribonucleic acid) in 1953 and subsequently, researchers throughout the world started working on the determination of DNA sequences. In 1977 two independent groups such as Maxam and Gilbert (1977) and Sanger, Nicklen, and Coulson (1977) were significant research groups who developed two different approaches of DNA sequencing (;). Initially Maxam–Gilbert method was preferred but later Sanger’s sequencing method gained more popularity for the development of efficient and faster sequencing technologies. The International Rice Genome Sequencing Project and *Arabidopsis* Genome Initiative and Human Genome Project were based on Sanger’s sequencing technology. Moreover, Sanger’s dideoxy sequencing technology is gold standard for genome sequencing, but it has several drawbacks such as in vivo cloning of DNA fragments, low throughput, time-consuming, high cost and require more labor (Sharma et al., 2017). To overcome these shortcomings, scientists and bioengineers have developed the new sequencing technologies that are called next-generation sequencing (NGS) or high-throughput sequencing (HTS) technologies.

Genome informatics involves identification of short stretch of DNA fragments to the sequencing of whole genome of an organism. Current progress in genome informatics provides comprehensive growth in lab benchwork as it is progressively reducing the hit-and-trial experiments (<https://www.ingentaconnect.com>). It comes out from the requirement of suitable informatics for management, distribution, and organization of biological data. It is an interdisciplinary field emerging from the interaction of molecular biology, statistics, mathematics, and computer science to study the genomic data of various organisms. It also predicts the structure and function of macromolecules (Aslam, Khattak, Ahmed, & Asif, 2017). Genome informatics explores the different aspects (genomics, transcriptomics, proteomics, metabolomics, metagenomics and epigenetics) of biomolecular organization, and complicated biological systems from cells to ecosystems. Due to the advancement of NGS technologies with low cost, there is a flood of molecular data which are generated from environmental samples to organisms (Esposito et al., 2016). The NGS data analysis is an important and essential technique for interpreting and analyzing vast amount of information generated and constructed using various biological approaches. With the aid of next-generation technologies, bioinformatics plays a vital role in the coding and decoding of genes, genomes, and proteins (Harishchander, 2017). The whole-genome sequencing of several species permits us to understand their structural and functional organization (Morrell, Buckler, & Ross-Ibarra, 2012; Weigel & Mott, 2009). Significant advancements in plant genomics have pushed the bioinformatics to a higher level than before in the field of agricultural research.

In addition, the sequencing of transcriptome and proteome plays an essential role in deciphering the content and functionality of gene(s) in genome organization (Chiusano, D’Agostino, Barone, Carputo, & Frusciante, 2009; Van Emon, 2015). Furthermore, the classification of the genes and genome and their interaction are vital to be deciphered into breeding training for livestock and crops, which contribute to their productivity, resistance, and health. The plant,

livestock, and soil microbiome also play major roles in agriculture to determine the soil's biogeochemical properties (Acosta-Martinez et al., 2014), plant fitness (Haney, Samuel, Bush, & Ausubel, 2015; Timmusk et al., 2014), quality, and yield traits (Babu, Jogaiah, Ito, Nagaraj, & Tran, 2015). The contribution of genomics and transcriptomics to agriculture spans the discovery and the manipulation of genes associated with breeding by marker-assisted selection (MAS) of variants (Iovene, Barone, Frusciante, Monti, & Carputo, 2004) as well as specific phenotypic traits (Zhang et al., 2014). This is so-called agrigenomics, which aims to find novel solutions through the study of crop genomics so that crop production can be further improved (Esposito et al., 2016). It maintains continuous productivity for the food or pharma industry, such as the design of pharmaceuticals or energy production (Blanchfield, 2004; Yuan, Tiller, Al-Ahmad, Stewart, & Stewart, 2008).

In 2005 NGS technology became commercially accessible. Roche/454 GS FLX + GS 20 sequencing technology was the first sequencing technology, and after that, many sequencing platforms and chemistries have been developed. These sequencing methods grouped into three different types, such as single-molecule sequencing, sequencing by ligation (SBL), and sequencing by synthesis (SBS). Till date, many NGS platforms have been developed, such as Illumina series, Ion Torrent, Life/AB SOLiD series, Helicos BioSciences: HeliScope, Pacific BioScience: PacBio, and Nanopore (Egan et al., 2012; Jain et al., 2018).

4.2 The evolution of DNA-seq

Maxam–Gilbert and Sanger sequencing were the main sequencing methods that used until the development of new sequencing technologies, which completely transformed the genome exploration and analysis approach (Maxam & Gilbert, 1977; Sanger et al., 1977). In 2005 454 Life Sciences discovered the first NGS technologies and commercialized by the Roche company, it was capable of generating huge amount of sequences with high speed, low cost, and reduced labor. These new sequencing technologies were further termed as “next-generation sequencing (NGS) technologies” (Qiang-long, Shi, Peng, & Fei-shi, 2014).

To date, various next NGS technologies have been developed that produce millions to billions of reads of many samples parallelly in a single run at much lower cost within an hour or a day. For example, Sanger's sequencing technology took approximately 15 years and 100 million US dollars to sequence human genome that contains approximately 3 billion bps distributed in 23 chromosomes which are located in each human cell nucleus, whereas 454 Genome Sequencer FLX took approximately 2 months at very low cost to sequence the human genome (Kchouk, Gibrat, & Elloumi, 2017; Mardis, 2011). After the invention of basic sequencing methods, many technologies have been developed, which can be categorized into three generations, that is, first, second, and third.

4.2.1 The first generation of sequencing technologies

Sanger and Maxam–Gilbert developed the DNA sequencing technologies in 1977, which are classified as the first-generation sequencing technologies (Heather & Chain, 2016; Liu et al., 2012). These technologies are based on different chemistries of DNA sequencing. Maxam–Gilbert approach based on chemical method (chemical degradation method) is dependent on the splitting of nucleotides by chemicals which is found to be more effective in smaller nucleotides. The chemical treatment breaks nucleotide into small proportion of bases into four reactions, that is, T + C, A + G, T, and G (Masoudi-Nejad, Narimani, & Hosseinkhan, 2013; Maxam & Gilbert, 1977). Due to its more complexity and low resolution, the Maxam and Gilbert approach did not gain much recognition. On the other side, Sanger's sequencing approach is based on dideoxynucleotide or chain termination method or SBS method (Sanger et al., 1977). It uses single strand of double-stranded DNA sequence as template for sequencing. This approach makes the use of chemical analogs of the deoxyribonucleotides (dNTP). Lacking 3' hydroxyl group in dideoxynucleotides (ddNTPs) which is necessary for the extension of DNA chain and due to which it cannot form a bond with 5' phosphate of the next dNTP. Subsequently, certain amounts of ddNTP labeled with radioactive isotopes are mixed (ddTTP, ddATP, ddGTP, and ddCTP) into four DNA extension reactions, respectively. Autoradiography and gel electrophoresis are used to determine the DNA sequence on the basis of their position (Chidgeavadze et al., 1984; Heather & Chain, 2016).

4.2.2 The second generation of sequencing technologies

Sanger sequencing method was dominant in the world for three decades. After that, in 2005 the emergence of NGS methods has revolutionized the whole-genome sequencing technologies. The second-generation sequencing technologies were produced by Roche/454 Life Sciences launched in 2005, Illumina in 2006, and ABI/SOLiD in 2007.

The main features of the second generation of sequencing technologies are that they generate millions of reads parallelly with high speed, low cost, and cheap labor, and output of reads can be detected without the need for electrophoresis. In this generation, it omitted the requirement of *in vivo* cloning, and DNA can be fragmented to generate sequencing libraries by using adapter ligation amplification using PCR (polymerase chain reaction)-based system. On the basis of chemistry, second-generation sequencing approaches are divided into two categories: SBL used by ABI's SOLiD and the second is SBS which is used by Illumina, Roche/454, and Ion Torrent (Heather & Chain, 2016; Sharma et al., 2017).

4.2.3 The third generation of sequencing technologies

The second-generation sequencing technology changes the traditional DNA sequencing approach, but it still requires PCR amplification step which takes long time, is complex protocol, and is very expensive in library preparation and sequencing. Moreover, due to repetitive regions in the genome, it becomes difficult and incapable of solving in second generation because of short reads. To overcome this issue, researchers developed the third-generation sequencing platforms such as PacBio and Oxford Nanopore. These new technologies generated the long reads (several kilobases) without the use of PCR amplification, cheaper in cost, and easy and straightforward sample preparation protocol as compared to previous generations (Goodwin, McPherson, & McCombie, 2016). Single-molecule real-time (SMRT) sequencing approach and synthetic approach are the two approaches used in third-generation sequencing technologies. SMRT was developed by Quake Laboratory, and synthetic approach was developed by Illumina (Moleculo) and 10xGenomics (Bentley, Balasubramanian, & Swerdlow, 2008; Braslavsky et al., 2003; Harris, Buzby, & Babcock, 2008). Both PacBio and Oxford Nanopore use the SMRT approach to generate sequencing data (Heather & Chain, 2016). Single-molecule sequencing platform has the ability to detect epigenetic modifications, and longer reads are also helpful in *de novo* genome assembly, improving previous genome assemblies with more accurate results. But the major drawback of single-molecule sequencing technology is its higher error rate (Sharma et al., 2017). The sequencing platform of each generation sequencing platforms and their details are listed in Table 4.1.

4.3 Genomics in agriculture

The NGS methods produce millions of sequencing reads which have not only simplified genome and transcriptome sequencing but also started to change the research in life sciences. This can be useful in whole-genome sequencing, structural variation discovery, simple sequence repeats (SSRs), single-nucleotide polymorphism (SNP), mRNA and non-coding RNA profiling, epigenomics, and chromosome chromatin conformation. These approaches provide useful information to solve the complex biological problems in agriculture, nutrition, and food (Esposito et al., 2016; Liu, 2009).

4.3.1 Genome assembly

De novo genome assembly is one of the major purposes of DNA sequencing. High-throughput genome sequencing provides practical solution to many challenges that occurred in the field of crop genomics such as novel *de novo* genome

TABLE 4.1 List of sequencing platforms as per generation.

Generations	Platform	Avg read length (bp)	Reads per run	Data generated per run
First	ABI Sanger	400–900*	96	0.00069–0.0021 Gb
Second	454	100–700	~1M	0.02–0.7 Gb
	Illumina	150–300	25M–6B	7.5 Gb–1.8 Tb
	SOLiD	75	3–6B	160–320 Gb
	Ion Torrent	200–400	0.4–80M	0.06–10 Gb
Third	PacBio	1300–13,500	350–600	0.5–7 Gb
	Oxford Nanopore	9545	100	1.5 Mb–4 Tb

B, Billion; *Gb*, gigabytes; *M*, million; *Tb*, terabytes.

assembly without the prior knowledge of reference genomic information. It refers to the method of taking millions of short DNA reads and putting back together to create chromosome wise distribution of sequences of an organism. Due to rapid growth in NGS technologies and the bioinformatics approaches that change the prospective of research from classical conservation genetics method to conservation genomics method (Allendorf, Hohenlohe, & Luikart, 2010; Primmer, 2009). Genome-wide data analysis provides detailed information of candidate gene approaches or genetic variations and this opens the scope of screening of selective variations and assessing the adaptive potential of populations such as quantitative trait loci (QTL) mapping, population selection and association mapping (Primmer, Papakostas, Leder, Davis, & Ragan, 2013; Steiner et al., 2013). Many large-scale data types such as RAD-seq, transcriptome sequencing, genotyping-by-sequencing, reduced representation sequencing, and amplicon sequencing can successfully utilize in many plant-based research without relying on a reference genome (Ekblom & Wolf, 2014). But a complete and annotated whole genome with chromosomal and positional data provides the crucial information for genomic approach such as SSR, SNP, InDels (insertion and deletions), CNV (copy number variation), and structural rearrangements, which is important for population-based genetic studies (Ellegren, Smeds, & Burri, 2012). All these studies strongly rely on complete and well-annotated genomic data for the identification of structural and functional genomic regions of interest (Ekblom & Wolf, 2014).

In 2005 Arabidopsis Genome Initiative published the genome of first plant species, that is, *Arabidopsis thaliana* and after that the genome of many plant species was got sequenced using first-generation and NGS technologies (Bevan & Walsh, 2005; Nivedita, Yadav, & Gautam, 2015; Turktas et al., 2015). There is a rapid growth in plant genome sequencing, which can provide deep knowledge of physiological and biochemical processes which showed how the plants respond to different environment factors such as abiotic and biotic stresses (Barba et al., 2014; Gedil, Ferguson, & Girma, 2016a). There are more than 100 plant genomes, which have already been decoded in past two decades such as *A. thaliana*, *Oryza sativa*, *Triticum aestivum*, *Glycine max*, *Zea mays*, *Brachypodium distachyon*, *Vigna radiata*, and *Vigna angularis*. Availability of these genomes has paved a new path for studying the evolutionary history, complex life cycle, phylogenetic relationship with other species, and functional and structural genomic organization. There are several nonfunctional model plant species such as *Utricularia gibba*, *Spirodela polyrhiza*, *Selaginella moellendorffii*, and *Genlisea aurea* which were sequenced. In addition, these plant species provide support in the evolutionary relationship between basal vascular plants and most diverse as well as complicated angiosperms (Sharma et al., 2017).

With the advancement in NGS technologies, it has become very affordable and feasible to assemble and annotate the genomic data of an organism. Decoding of high-quality genome assembly of eukaryotes is a very challenging problem that requires high computational resources, software, expertise, and time due to its complex genomic structure, repeats, and sizes (Badouin, Gouzy, & Grassa, 2017; Jansen, Liem, & Jong-Raadsen, 2017). Fungi, virus, and bacterial genomes require less resources and time as compared to eukaryotes. Eukaryote genomes take weeks to months for running a genome assembly and annotation (Dominguez Del Angel, Hjerde, & Sterck, 2018).

4.3.1.1 Pipeline of genome assembly

Every genome assembly project is different and complex on the basis of species to species (Fig. 4.1). There are several properties that are needed to know before going to start genome assembly projects of an organism on the basis of complexity and properties (Dominguez Del Angel et al., 2018) such as:

1. estimated genome size, repeats, heterozygosity, ploidy level, and GC content;
2. high-quality DNA requirements for de novo sequencing, chemical purity, and structural integrity of DNA;
3. suitable sequencing technology (short- or long-read sequencing technology);
4. computational resources;
5. genome assembly;
6. transposable elements;
7. structural and functional annotation;
8. high-quality genome assembly submission.

Steps involved in genome assembly are as follows:

Step 1. Preprocessing of data: In this step, to check the quality of fastq data and removal of low-quality reads, very short reads, adapters, and overrepresentative sequences, reads with phred score less than 20, N's in reads, are carried out. Trimmomatic (Bolger, Lohse, & Usadel, 2014) and Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) can be used for the removal of low-quality data.

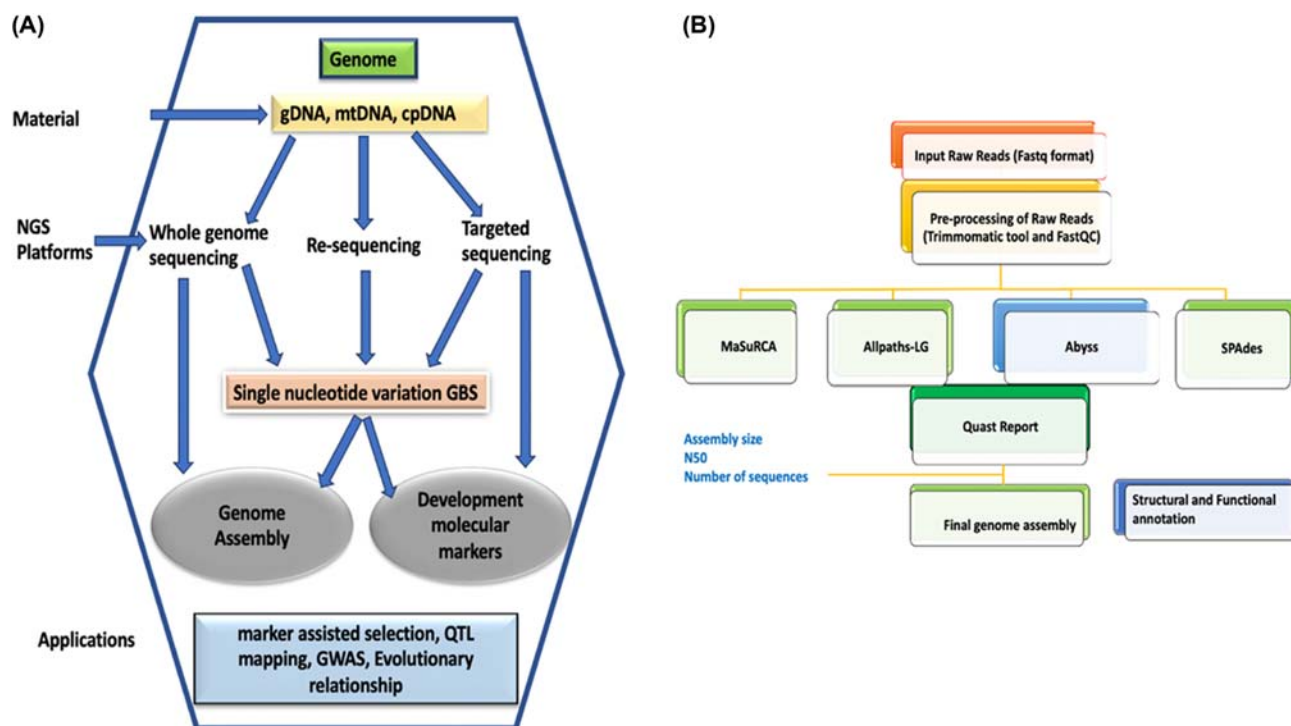


FIGURE 4.1 (A) Usage of NGS platform for genome sequencing, assembly, and analysis; (B) pipeline of genome assembly analysis. NGS, Next-generation sequencing.

Step 2. Genome assembly: KmerGenie tool is used to estimate the best k-mer length size of genome and on the basis of estimated k-mers, de novo assembly is to run by using short-read or long-read assembly tool such as MaSuRCA (Zimin, Marçais, & Puiu, 2013), Allpaths-LG (MacCallum, Przybylski, & Gnerre, 2009), Abyss (Simpson et al., 2009), SPAdes (Bankevich et al., 2012), and FALCON (Chin et al., 2016).

Step 3. Quality of assembly: Quality of de novo assembly on the basis of N50, GC content, and number of sequences is checked by QUAST server (Gurevich, Saveliev, & Vyahhi, 2013).

Step 4. Downstream analysis: After getting the final de novo genome assembly, the next step is finding the exon and intron organization, genic region identification, structure and functional annotation, repeats and variants identification, motifs and domains, transcriptional factors, etc. (Fig. 4.2).

4.3.1.2 Simple sequence repeats

Microsatellites also called SSRs are the codominant markers that play a crucial role in biological functions. It is present in the form of DNA sequences which contain 1–6 repeating units of nucleotides (Kapil, Rai, & Shanker, 2014). These repeating units of nucleotides are present in both intro and exon region of DNA, but it is more abundant in the intron region (Yu, Dossa, & Wang, 2016). Microsatellite markers play a crucial role in association studies, linkage mapping, MAS, diversity evaluation, gene mapping, fingerprinting, and species identification for crop improvement due to its codominant inheritance, multiallelic nature, robust amplification, and reproducibility nature. Microsatellite becomes one of the most powerful and vital techniques for plant genetic studies (Wang, Elbaidouri, & Abernathy, 2015; Wang, Do Kim, & Gao, 2016). Microsatellite markers are extremely polymorphic as it depends on an amount of short tandem repeats (Chen, Liu, & Wang, 2015).

Microsatellite discovery using traditional methods from genomic data is compromised with a number of markers besides cost, labor, and time. To overcome this issue, an in silico approach is another alternative. The in silico approach has the benefit of predicting target-specific region in genome which can be more effective in developing molecular markers required for linkage mapping and QTL (Gupta, Souframanien, & Gopalakrishna, 2008; Sharma et al., 2017). Linkage map requires more molecular markers to increase the map density for fine mapping (Marubodee, Ogiso-Tanaka, & Isemura, 2015). MISA (MICROSATellite identification tool), a widely used Perl script for mining SSRs, has certain set

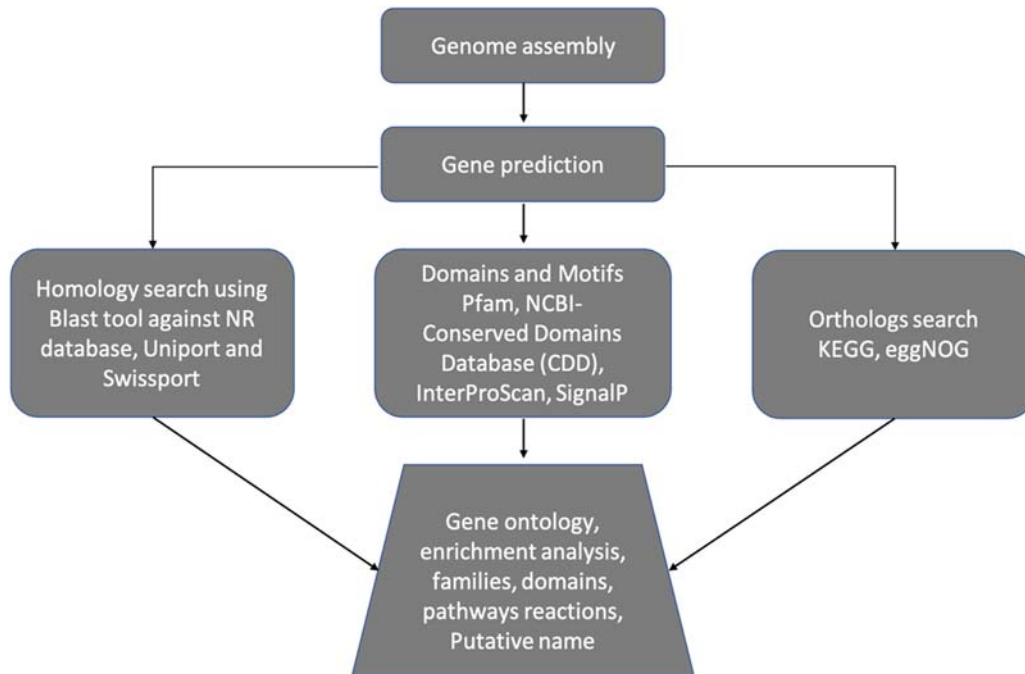


FIGURE 4.2 Workflow of gene prediction and annotation.

of default parameters, that is, 10 repeating units for mono-nucleotide, 6 units for di-nucleotide, and 5 repeating units for tri-nucleotide, tetra-nucleotide, penta-nucleotide, and hexa-nucleotide (Thiel, Michalek, & Varshney, 2003).

4.3.2 RNA-seq in agriculture

Transcriptome is the protein-coding part of genomic data of an organism and it refers to the set of RNA molecules such as mRNA (messenger RNA), tRNA (transfer RNA), rRNA (ribosomal RNA), and ncRNA (noncoding RNA) which are present in the cells. Whole transcriptome shotgun sequencing or RNA-seq is the study of whole set of RNAs transcribed in a cell and their quantity for certain physiological condition and developmental stage (Gedil, Ferguson, & Girma, 2016b).

RNA-seq uses the NGS technology to identify the quantity and presence RNA in a biological sample at a given time point. This is found to be more accurate and sensitive way to study the genome-wide differential expressed genes and it overcomes the limitation of microarray (Voelckel, Gruenheit, & Lockhart, 2017). RNA-seq revolutionized the field of agriculture transcriptomics where genomes of nonmodel species are rarely available (Fig. 4.3).

There are many studies carried out on plant, since the decoding of *Arabidopsis* genome. The RNA-seq analysis was successfully applied in rice, wheat, maize, lentils, mung bean, and so on. RNA-seq profiling of *Brassica napus* was performed to identify various biochemical pathways (Sharma et al., 2017). Transcriptome analysis of root tissue of wheat was carried out, which reported 45,139 differential expressed genes (DEGs) in four sets of control and treated samples of resistant and susceptible cultivars. Furthermore, 13,820 TF and 435,829 genic putative markers were identified (Iqbal, Sharma, & Jasrotia, 2019). A total of 6310 DEGs were found to be associated with herbicide tolerance (Iqbal, Soren, & Gangwar, 2017).

4.3.2.1 Types and pipeline of RNA-seq

RNA-seq analysis is categorized into reference-based and de novo-based methods (Fig. 4.4).

- 1. Reference-based RNA-seq analysis:** This approach is used when reference genome assembly and its annotation is already available.
- 2. De novo-based RNA-seq analysis:** This approach is used when genome of specific species is not available. In this case, all the reads of each biological replicates of various conditions are pooled, and the de novo transcriptome assembly is carried out.

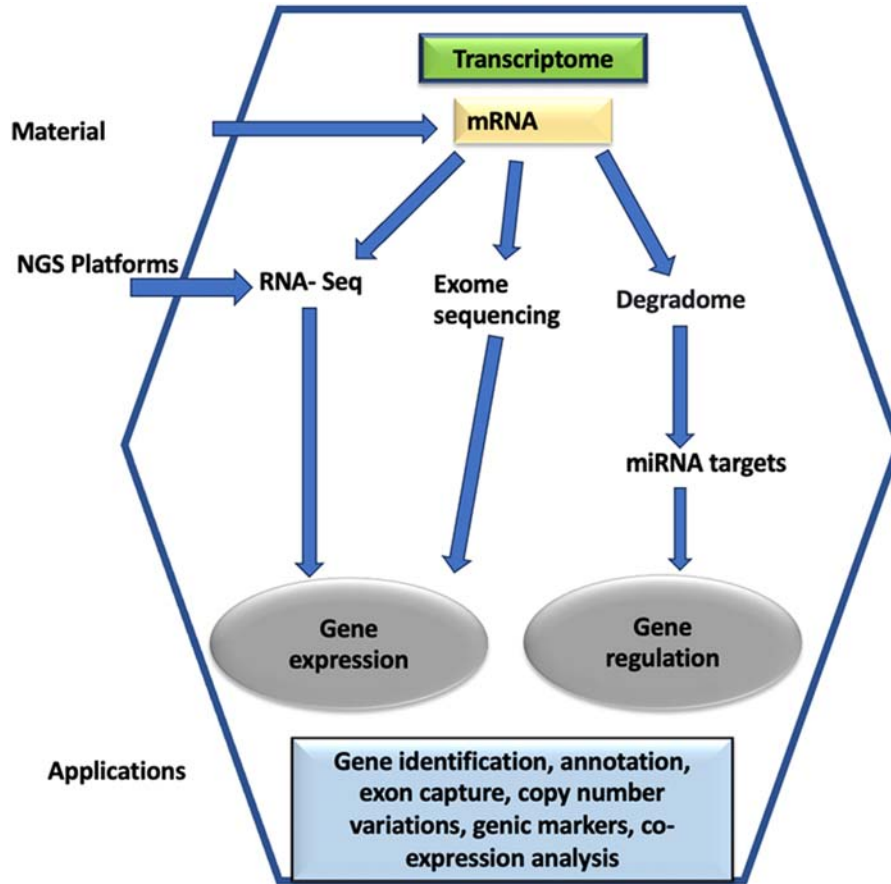


FIGURE 4.3 Usage and application of NGS technologies in transcriptome analysis. NGS, Next-generation sequencing.

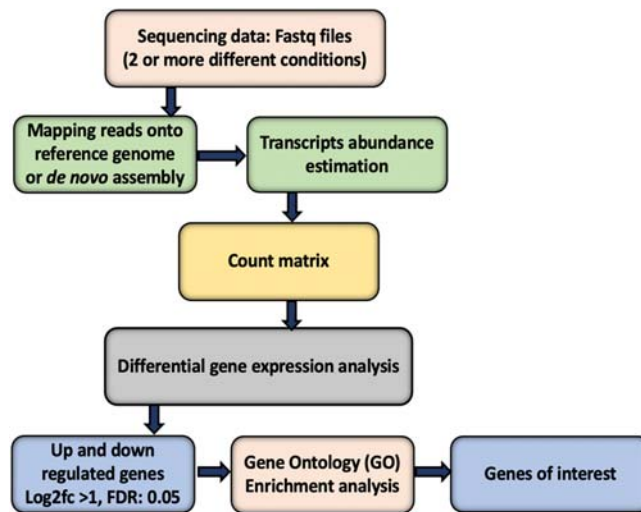


FIGURE 4.4 Pipeline of RNA-Seq analysis.

TABLE 4.2 List of software and tools for next-generation data analysis.

Category	Tool name	Aims and scope
Preprocessing	FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)	Assessment of reads
	FASTX-toolkit (Gordon and Hannon (2010), NGS QC Toolkit (Patel and Jain, 2012), Cutadapt (Martin, 2011), Trimmomatic (Bolger et al., 2014)	Removal of low-quality reads and adaptors
Assembly	ALLPATHS (Butler et al., 2008), SOAPdenovo-Trans (Xie et al., 2014), Trans-ABYSS (Robertson et al., 2010), Oases (Schulz et al., 2012), Spades (Bankevich et al., 2012), SOAP-denovo2 (Luo et al., 2012), ABYSS (Simpson et al., 2009), Velvet (Zerbino and Birney, 2008), Trinity (Haas et al., 2013), FALCON (Chin et al., 2016), MIRA (Chevreux et al., 1999), rnaSPAdes (Bushmanova et al., 2019)	Genome and transcriptome assembly tools
Mapping	Tophat2 (Kim et al., 2013), STAR aligner (Dobin et al., 2013), Bowtie2 (Langmead and Salzberg, 2012), HISAT (Kim et al., 2015), BWA (Li and Durbin, 2009)	Read aligner tools
Expression analysis	RSEM (Li and Dewey, 2011), Feature counts (Liao et al., 2014), HTSeq (Anders et al., 2015), Kallisto (Bray et al., 2016)	RNA-seq read count tools
Differential expression analysis	EdgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014), NOISeq (Tarazona et al., 2015), EBSeq (Leng et al., 2013)	R package for differential expression analysis
Single nucleotide polymorphism and population genomics	SAMtools (Li et al., 2009), Plink (Purcell et al., 2007), Tassel (Bradbury et al., 2007), GATK (Van der Auwera et al., 2013), VcfTools (Danecek et al., 2011), Stacks (Catchen et al., 2013)	SNP and GWAS analysis
Markers	MISA (Beier et al., 2017), GMATo (Wang et al., 2013)	Simple sequence repeats tools

Steps involved in analysis are as follows:

Step 1: Preprocessing of data: It includes removal of low-quality reads on the basis of phred score less than 20, very short reads, adaptors and overrepresentative sequences, N's in the reads, etc.

Step 2: Mapping: In this step, all the reads of each condition separately map onto reference genome or de novo transcriptome assembly. There are various tools developed for the alignment of reads onto reference genome such as Hisat2 (Kim, Paggi, Park, Bennett, & Salzberg, 2019), Tophat2 (Kim, Pertea, & Trapnell, 2013), STAR (Dobin, Davis, & Schlesinger, 2013), and Bowtie2 (Langmead & Salzberg, 2012).

Step 3: Transcript abundance estimation: Abundance estimation and read count can be obtained by using RSEM (Li & Dewey, 2011), featureCounts (Liao, Smyth, & Shi, 2014), or HTSeq (Anders, Pyl, & Huber, 2015).

Step 4: Differential expression analysis: Expression analysis of genes can be detected from two different groups at specific time point or other conditions such as control and treated, etc. Several tools, such as EdgeR (Robinson, McCarthy, & Smyth, 2010), DESeq2 (Love, Huber, & Anders, 2014), and NOIseq (Tarazona, Furió-Tarí, & Turrà, 2015), can be used for differential expression analysis.

Step 5: Homology search and gene ontology: After getting the DEGs, next step is finding the gene name based on homology search against NCBI's NR database (NRDB) and subsequently gene ontology can be analyzed using the blast2GO tool (Conesa, Götz, & García-Gómez, 2005).

The list of software and tools used for NGS data analysis are summarized in Table 4.2, which also include genome assembly, RNA-seq, and SSR analysis.

4.3.3 Databases and prediction servers

Storage of NGS data as well as analyzed data is also a substantial issue to scientific community. Several biological databases have been developed to overcome this issue, and these databases are NCBI (<https://www.ncbi.nlm.nih.gov/>), DDBJ (<https://www.ddbj.nig.ac.jp/index-e.html>), and EMBL (<https://www.embl.org/>). All three databases share and

TABLE 4.3 List of databases for storing next-generation sequencing data.

Data type	NCBI	DDBJ	EMBL-EBI
Next-generation sequencing reads	SRA	DRA	ENA
Capillary reads	Trace Archive	DTA	
Annotated sequences	GenBank	DDBJ	

DRA, DDBJ Sequence Read Archive; *DTA*, DDBJ Trace Archive; *ENA*, European Nucleotide Archive; *SRA*, Sequence Read Archive.

TABLE 4.4 List of important plant specific databases.

Rice Genome Annotation Project	<i>Oryza sativa</i>	http://rice.plantbiology.msu.edu/
TAIR	<i>Arabidopsis thaliana</i>	https://www.arabidopsis.org/
PlantGDB	Multiple plants species	http://www.plantgdb.org/AtGDB/
Phytozome	Multiple plants species	https://phytozome.jgi.doe.gov/pz/portal.html
EnsemblPlantys	Multiple plants species	https://plants.ensembl.org/index.html
PGDBj	Multiple plants species	http://pgdbj.jp/?ln=en
Sol Genomics Network	Multiple plants species	https://solgenomics.net/
Pulse crop database	List of pulses only	https://www.pulsedb.org/
PlantTFDB	Plant transcriptional factor database	http://planttfdb.gao-lab.org/

PGDBj, The Plant Genome DataBase Japan; *TAIR*, The Arabidopsis Information Resource.

update their data with each other regularly through the International Nucleotide Sequence Database Collaboration (<http://www.insdc.org/>). The most important databases which share NGS data with each other are listed in Table 4.3.

Furthermore, there are several other databases which are useful for NGS databases such as Gene Expression Omnibus (repository of microarrays data and high-throughput gene expression data) (<https://www.ncbi.nlm.nih.gov/geo/>), Transcriptome Shotgun Assembly Sequence Database (for transcriptome assembly) (<https://www.ncbi.nlm.nih.gov/genbank/tsa/>), ArrayExpress (<https://www.ebi.ac.uk/arrayexpress/>), and Ensembl (<https://ensembl.org/index.htm>).

4.3.3.1 List of plant-specific databases

There are specific databases that have been developed for various plant organisms which contain the information of genome assembly, gene annotation, markers, variants, protein sequences, motifs, domains, etc. (Table 4.4).

4.4 Conclusion, applications, and future prospects of next-generation sequencing in agriculture

NGS technology plays a vital role in transforming the experimental design at molecular level, which enables us in increasing scientific knowledge in the field of agricultural research. Applications of NGS technologies have significantly speed up the whole-genome sequencing projects and resequencing of genome to RNA-Seq, DNA methylation sequencing, and epigenomics and metagenomics. It plays an important role in structural and functional genomics, plant breeding, ecology, evolution, and plant disease diagnosis. Furthermore, markers including SNPs and SSRs are the most suitable predominant marker types exploited in plant breeding approaches (Egan et al., 2012).

Genome assembly, transcriptomics, metagenomics, and epigenomics may also contribute to the understanding of the functionality and organization of biological systems that provide insight to trace the molecular variability during the development stage under specific condition such as pathological (biotic) or influenced by environmental and/or

physiological changes (abiotic) (Esposito et al., 2016). Genome assemblies provide an opportunity to mine millions of molecular markers such as SNPs and SSRs, and characterization of agronomically important genes. It has been observed that variants such as SNPs and InDels dominate the molecular markers' applications and usage due to their progression in NGS technology (Edwards & Batley, 2010). Molecular markers play an important role in the development of genetic and physical maps of genome and also involve in finding genes or QTL, which helps in regulating economically key traits (Varshney, Nayak, & May, 2009).

Postadvancement of HTS technologies, large-scale genome-wide study of evolutionary, and phylogenetic and comparative analysis has become much easier for model and nonmodel crops (Grover, Salmon, & Wendel, 2012). HTS has revolutionized the field of agricultural genomics research (agrigenomics). The genome assembly of organism, transcriptomics, epigenomics, and metagenomics are the major areas which are influenced by NGS technology. Continuous advancement in the NGS technologies tends to decrease the cost, labor, and time, which has opened the door for small laboratories as well as researchers.

References

- Acosta-Martinez, V., Cotton, J., Gardner, T., Moore-Kucera, J., Zak, J., Wester, D., . . . Cox, S. (2014). Predominant bacterial and fungal assemblages in agricultural soils during a record drought/heat wave and linkages to enzyme activities of biogeochemical cycling. *Applied Soil Ecology*, *84*, 69–82.
- Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, *11*(10), 697–709.
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169.
- Aslam, Z., Khattak, J. Z. K., Ahmed, M., & Asif, M. (2017). A role of bioinformatics in agriculture. In M. Ahmed, & C. Stockle (Eds.), *Quantification of climate variability, adaptation, and mitigation for agricultural sustainability* (pp. 413–434). Cham: Springer.
- Babu, A. N., Jogaiah, S., Ito, S. I., Nagaraj, A. K., & Tran, L. S. P. (2015). Improvement of growth, fruit weight and early blight disease protection of tomato plants by rhizosphere bacteria is correlated with their beneficial traits and induced biosynthesis of antioxidant peroxidase and polyphenol oxidase. *Plant Science (Shannon, Ireland)*, *231*, 62–73.
- Badouin, B., Gouzy, J., Grassa, C. J., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, *546*(7656), 148–152.
- Bankevich, A., Nurk, S., Antipov, D., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477.
- Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, *6*(1), 106–136.
- Beier, S., Thiel, T., Münch, T., et al. (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics (Oxford, England)*, *33*(16), 2583–2585.
- Bentley, D., Balasubramanian, S., Swerdlow, H., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*, 53–59.
- Bevan, M., & Walsh, S. (2005). The Arabidopsis genome: A foundation for plant research. *Genome Research*, *15*(12), 1632–1642.
- Blanchfield, J. R. (2004). Genetically modified food crops and their contribution to human nutrition and food quality. *Journal of Food Science*, *69*(1), CRH28–CRH30.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120.
- Bradbury, P. J., Zhang, Z., Kroon, D. E., et al. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635.
- Braslavsky, I., Hebert, B., Kartalov, E., & Stephen, R. (2003). Quake Sequence information can be obtained from single DNA molecules. *PNAS*, *100*(7), 3964.
- Bray, N. L., Pimentel, H., Melsted, P., et al. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.
- Bushmanova, E., Antipov, D., Lapidus, A., et al. (2019). rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience*, *8*(9), giz100.
- Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., . . . Jaffe, D. B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Research*, *18*(5), 810–820.
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.
- Chen, H., Liu, L., Wang, L., et al. (2015). Development and validation of EST-SSR markers from the transcriptome of adzuki bean (*Vigna angularis*). *PLoS One*, *10*(7), e0131939.
- Chevreur, B., Wetter, T., & Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. *German conference on bioinformatics*, *99*(1), 45–56.

- Chidgeavazde, Z. G., Beabealashvilli, R. S., Atrazhev, A. M., Kukhanova, M. K., Azhayev, A. V., & Kravetsky, A. A. (1984). 2',3'-Dideoxy-3' aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Research*, *12*(3), 1671–1686.
- Chin, C. S., Peluso, P., Sedlazeck, F. J., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*(12), 1050–1054.
- Chiusano, M. L., D'Agostino, N., Barone, A., Carputo, D., & Frusciante, L. (2009). Genome analysis of species of agricultural interest. In P. J. Papajorgji, & P. M. Pardalos (Eds.), *Advances in modeling agricultural systems* (25, pp. 385–402). Boston, MA: Springer.
- Conesa, A., Götz, S., García-Gómez, J. M., et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*(18), 3674–3676.
- Danecek, P., Auton, A., Abecasis, G., et al. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158.
- Dobin, A., Davis, C. A., Schlesinger, F., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* *2013*, *29*(1), 15–21.
- Dominguez Del Angel, V., Hjerde, E., Sterck, L., et al. (2018). Ten steps to get started in genome assembly and annotation. *F1000Res*, *7*, ELIXIR-148.
- Edwards, D., & Batley, J. (2010). Plant genome sequencing: Applications for crop improvement. *Plant Biotechnology Journal*, *8*(1), 2–9.
- Egan, A. N., Schlueter, J., & Spooner, D. M. (2012). Applications of next-generation sequencing in plant biology. *American Journal of Botany*, *99*(2), 175–185.
- Eklblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042.
- Ellegren, H., Smeds, L., Burri, R., et al. (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, *491*(7426), 756–760.
- Esposito, A., Colantuono, C., Ruggieri, V., & Chiusano, M. L. (2016). Bioinformatics for agriculture in the next-generation sequencing era. *Chem Biol Technol Agric*, *3*(1), 9.
- Fadiel, A., Anidi, I., & Eichenbaum, K. D. (2005). Farm animal genomics and informatics: An update. *Nucleic Acids Research*, *33*(19), 6308–6318.
- Gedil, M., Ferguson, M., Girma, G., et al. (2016a). *Perspectives on the application of next-generation sequencing to the improvement of Africa's staple food crops*. *Next generation sequencing: Advances, applications and challenges* (10, pp. 287–321). InTech.
- Gedil, M., Ferguson, M., Girma, G., et al. (2016b). *Perspectives on the application of next-generation sequencing to the improvement of Africa's staple food crops*. *Next generation sequencing: Advances, applications and challenges* (pp. 2218–2248). InTech.
- Goodwin, S., McPherson, J., & McCombie, W. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*, 333–351.
- Gordon, A., & Hannon, G. J. (2010). *Fastx-toolkit. FASTQA short-reads pre-processing tools (unpublished)*. http://hannonlab.cshl.edu/fastx_toolkit/.
- Grover, C. E., Salmon, A., & Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany*, *99*(2), 312–319.
- Gupta, S. K., Souframanien, J., & Gopalakrishna, T. (2008). Construction of a genetic linkage map of black gram, *Vigna mungo* (L.) Hepper, based on molecular markers and comparative studies. *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, *51*(8), 628–637.
- Gurevich, A., Saveliev, V., Vyahhi, N., et al. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–1075.
- Haas, B. J., Papanicolaou, A., Yassour, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512.
- Haney, C. H., Samuel, B. S., Bush, J., & Ausubel, F. M. (2015). Associations with rhizosphere bacteria can confer an adaptive advantage to plants. *Nature Plants*, *1*(6), 15051.
- Harishchander, A. (2017). A review on application of bioinformatics in medicinal plant research. *Bioinformatics & Proteomics Open Access Journal*, *1*, 000104.
- Harris, T. D., Buzby, P. R., Babcock, H., et al. (2008). Single-molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, *320*(5872), 106–109.
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8.
- Iovene, M., Barone, A., Frusciante, L., Monti, L., & Carputo, D. (2004). Selection for aneuploid potato hybrids combining a low wild genome content and resistance traits from *Solanum commersonii*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *109*(6), 1139–1146.
- Iqbal, M. A., Sharma, P., Jasrotia, R. S., et al. (2019). RNAseq analysis reveals drought-responsive molecular pathways with candidate genes and putative molecular markers in root tissue of wheat. *Scientific Reports*, *9*(1), 13917.
- Iqbal, M. A., Soren, K. R., Gangwar, P., et al. (2017). Discovery of putative herbicide resistance genes and its regulatory network in chickpea using transcriptome sequencing. *Frontiers in Plant Science*, *8*, 958.
- Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., ... Malla, S. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338.
- Jansen, H. J., Liem, M., Jong-Raadsen, S. A., et al. (2017). Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports*, *7*(1), 7213.
- Kapil, A., Rai, P. K., & Shanker, A. (2014). ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants. *Database (Oxford)*, bau107.

- Kchouk, M., Gibrat, J. F., & Elloumi, M. (2017). Generations of sequencing technologies: From first to next generation. *Biology and Medicine (Aligarh)*, 9(3), 1–8.
- Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4), 357–360.
- Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8), 907–915.
- Kim, D., Pertea, G., Trapnell, C., et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), R36.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357.
- Leng, N., Dawson, J. A., Thomson, J. A., et al. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England)*, 29(8), 1035–1043.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079.
- Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.
- Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7), 923–930.
- Liu, G. E. (2009). Applications and case studies of the next-generation sequencing technologies in food, nutrition and agriculture. *Recent Patents on Food, Nutrition & Agriculture*, 1(1), 75–79.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., ... Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International*, 2012, 11.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
- Luo, R., Liu, B., Xie, Y., et al. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1), 18.
- MacCallum, I., Przybylski, D., Gnerre, S., et al. (2009). ALLPATHS 2: Small genomes assembled accurately and with high continuity from short paired reads. *Genome Biology*, 10, R103.
- Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470, 198–203.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17(1), 10–12.
- Marubodee, R., Ogiso-Tanaka, E., Isemura, T., et al. (2015). Construction of an SSR and RAD-marker based molecular linkage map of *Vigna vexillata* (L.) A. Rich. *PLoS One*, 10(9), e0138942.
- Masoudi-Nejad, A., Narimani, Z., & Hosseinkhan, N. (2013). *Next generation sequencing and sequence assembly: Methodologies and algorithms* (1st Ed.), p. 86)New York: Springer-Verlag, X.
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), 560–564.
- Morrell, P. L., Buckler, E. S., & Ross-Ibarra, J. (2012). Crop genomics: advances and applications. *Nature Reviews Genetics*, 13(2), 85.
- Nivedita, Yadav, P. K., & Gautam, B. (2015). Gene expression profiling of transcription factors of *Arabidopsis thaliana* using microarray data analysis. *International journal of advanced research in computer science and software engineering*, 5(4), 783–793.
- Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, 7(2), e30619.
- Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575.
- Primmer, C. R., Papakostas, S., Leder, E. H., Davis, M. J., & Ragan, M. A. (2013). Annotated genes and nonannotated genomes: Cross-species use of gene ontology in ecology and evolution research. *Molecular Ecology*, 22(12), 3216–3241.
- Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, 1162, 357–368.
- Qiang-long, Z., Shi, L., Peng, G., & Fei-shi, L. (2014). High-throughput sequencing technology and its application. *Journal of Northeast Agricultural University*, 21(3), 84–96.
- Robertson, G., Schein, J., Chiu, R., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11), 909–912.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–140.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–5467.
- Schulz, M. H., Zerbino, D. R., Vingron, M., et al. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28(8), 1086–1092.
- Sharma, T. R., Devanna, B. N., Kiran, K., Singh, P. K., Arora, K., Jain, P., ... Singh, J. (2017). Status and prospects of next generation sequencing technologies in crop plants. *Current Issues in Molecular Biology*, 27, 1–36.
- Simpson, J. T., Wong, K., Jackman, S. D., et al. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), 1117–1123.

- Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., et al. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, 1, 261–281.
- Tarazona, S., Furió-Tarí, P., Turrà, D., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21), e140.
- Thiel, T., Michalek, W., Varshney, R., et al. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 106(3), 411–422.
- Timmusk, S., El-Daim, I. A. A., Copolovici, L., Tanilas, T., Kännaste, A., Behers, L., & Niinemets, Ü. (2014). Drought-tolerance of wheat improved by rhizosphere bacteria from harsh environments: Enhanced biomass production and reduced emissions of stress volatiles. *PLoS One*, 9(5), e96086.
- Turktas, K., Kurtoglu, K. Y., Dorado, G., Zhang, B., Hernandez, P., & Unver, T. (2015). Sequencing of plant genomes? A review. *Turkish Journal of Agriculture and Forestry*, 39, 361–376.
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics/Editorial Board, Andreas D. Baxevanis... [et al.]*, 43(1110), 11.10.1–11.10.33.
- Van Emon, J. M. (2015). The omics revolution in agricultural research. *Journal of Agricultural and Food Chemistry*, 64(1), 36–44.
- Varshney, R. K., Nayak, S. N., May, G. D., et al. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, 27(9), 522–530.
- Voelckel, C., Gruenheit, N., & Lockhart, P. (2017). Evolutionary transcriptomics and proteomics: Insight into plant adaptation. *Trends in Plant Science*, 22(6), 462–471.
- Wang, X., Lu, P., & Luo, Z. (2013). GMATo: A novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformatics*, 9(10), 541–544.
- Wang, L. X., Elbaidouri, M., Abernathy, B., et al. (2015). Distribution and analysis of SSR in mung bean (*Vigna radiata* L.) genome based on an SSR-enriched library. *Molecular Breeding*, 35(1), 25.
- Wang, L., Do Kim, K., Gao, D., et al. (2016). Analysis of simple sequence repeats in rice bean (*Vigna umbellata*) using an SSR-enriched library. *Crop Journal*, 4(1), 40–47.
- Watson, J. D., & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–738.
- Weigel, D., & Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biology*, 10(5), 107.
- Xie, Y., Wu, G., Tang, J., et al. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics (Oxford, England)*, 30(12), 1660–1666.
- Yu, J., Dossa, K., Wang, L., et al. (2016). PMDBase: a database for studying microsatellite DNA and marker development in plants. *Nucleic Acids Research*, 45(D1), D1046–D1053.
- Yuan, J. S., Tiller, K. H., Al-Ahmad, H., Stewart, N. R., & Stewart, C. N., Jr (2008). Plants to power: bioenergy to fuel the future. *Trends in Plant Science*, 13(8), 421–429.
- Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821–829.
- Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., & Simianer, H. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS One*, 9(3), e93017.
- Zimin, A. V., Marçais, G., Puiu, D., et al. (2013). The MaSuRCA genome assembler. *Bioinformatics (Oxford, England)*, 29(21), 2669–2677.

This page intentionally left blank

Genomics and its role in crop improvement

Ujjawal Kumar Singh Kushwaha¹, Nav Raj Adhikari², Birendra Prasad³, Suresh Kumar Maurya⁴, Devarajan Thangadurai⁵ and Jeyabalan Sangeetha⁶

¹National Plant Breeding and Genetics Research Center, Nepal Agricultural Research Council, Khumaltar, Lalitpur, Nepal, ²Institute of Agriculture and Animal Science, Tribhuvan University, Kirtipur, Kathmandu, Nepal, ³Department of Genetics and Plant Breeding, College of Agriculture, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India, ⁴Department of Vegetable Science, College of Agriculture, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India, ⁵Department of Botany, Karnatak University, Dharwad, Karnataka, India, ⁶Department of Environmental Science, Central University of Kerala, Periyar, Kerala, India

5.1 Introduction

Tom Roderick, a geneticist of Jackson Laboratory, coined the word genomics on the mapping of human genome in 1986 but the term was used first time in 1926 (Mckusick, 2005; Ogbe, Ochalefu, & Olaniru, 2016). “Genome” is composed of two independent words gene and omics where gene (Greek) refers to creation or birth and omics means the study of respective fields. Gene is a set of DNA which can synthesize protein independently and omics is the collective characterization and quantification of biological molecules which translate into structure, function, and dynamics of an organism. Therefore genome is the total genetic materials of an organism and genomics is the study of the total genes of an organism where total gene refers to whole genome of that organism.

Genomics could also be defined as the analysis of an organism complete DNA sequence (Bustamante, De La Vega, & Burchard, 2011; Hardison, 2003). The suffix omics when added to protein, genome, transcriptome, and metabolome make a sense and study of respective fields such as proteomics, genomics, transcriptomics, and metabolomics. Genomics is the study of whole genome of an organism which uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and functions of genomes. Most importantly, genomics focuses on interactions between loci and alleles within the genomes. It also deals with other interactions like epistasis, pleiotropy, and heterosis (Fig. 5.1) (Ogbe et al., 2016; Cooper, Kaufman, & Ward, 2003). The work on genomics and next-generation sequencing technology is pioneered by Fred Sanger. Fred Sanger and his colleagues established the technique of sequencing, genome mapping, data storage, and bioinformatics analysis between 1970 and 1980. This work paved the establishment of Human genome Project in 1990 and completed with the publication of complete human genome sequence in 2003 (<http://www.ebi.ac.uk/>).

5.1.1 Genome

Genomics is the study of whole genes of an individual, its interactions with each other, and the environments (<http://www.genome.gov>). It deals with the structure, function, evolution, mapping, and editing of genomes. A genome is the complete set of DNA of an organism (Hardison, 2003; Ogbe et al., 2016). Every single cell in the human body contains nearly 3 billion DNA base pairs that make up the human genome (Lander, 1996). A DNA contains four language letters, adenine, thymine, cytosine, and guanine which hold the information needed to build the entire human body. Similarly, a gene is a unit of DNA that directs the synthesis of a specific protein. A human body contains nearly 20,000–25,000 genes and each gene codes for an average of three proteins. These genes are located in 23 pairs of chromosomes in a nucleus of human cell. The main function of these genes is to produce proteins with the assistance of enzymes and messenger molecules. The role of enzyme is to copy the information from gene's DNA into a molecule

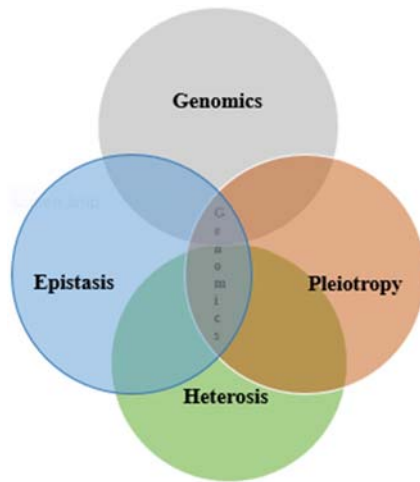


FIGURE 5.1 Genomic studies the interactions like epistasis, pleiotropy, and heterosis.

called messenger RNA (mRNA). The mRNA travels out of the nucleus into the cell cytoplasm where the mRNA is read by a ribosome and the information is used to link together small molecules, the amino acids to form a specific protein (<http://www.genome.gov>). Proteins make up the body organs, tissues and also control chemical reactions. If DNA mutates, it might produce abnormal protein and disturb body normal functioning (Cooper et al., 2003; Daly, Rioux, Schaffner, Hudson, & Lander, 2001; Lander, 1996).

The genome sequence of an organism is useful to understand the function of individual genes and their network for explaining evolutionary relationships and processes, and for depicting unknown regulatory mechanisms that coordinate the activities of genes (Bevan & Uauy, 2013). The genome-based approaches have multiapplications, for the disease diagnosis and treatments and for the improvement of food crops and fuel production. Genomics is an emerging and complex subject that can be categorized into different groups to make the study easy.

1. *Cognitive genomics*: It is the branch of genomics where we study the change in cognitive processes associated with genetic profiles.
2. *Comparative genomics*: It is the branch of genomics where we study structures and functions of different genomes of several biological species and relate them for comparative study. The comparison might be between DNA sequence, genes, gene order, regulatory sequence, and other genomic structures (Hardison, 2003).
3. *Functional genomics*: It is the branch of genomics where we study the function of genes, proteins, and their interactions. Functional genomics is often termed transcriptomics. Functional genomics makes use of vast data generated from genomics and transcriptomics projects (Schuler & Reichhart, 2003).
4. *Metagenomics*: It is the branch of genomics where we study the genetic materials recovered directly from environmental samples. Metagenomics is also referred to as environmental genomics, ecogenomics, or community genomics.
5. *Neurogenomics*: It is the branch of genomics that deals with the study of genetic materials that influences the structure and function of nervous system.
6. *Pangenomics*: It is the branch of genomics where we study the entire collection of genes and genomes found within a given species. The pangenome includes the core genome containing genes present in all strains within the clade.
7. *Personal genomics*: It is the branch of genomics which is concerned with the sequencing and analysis of the genome of an individual. Personal genomics is also called consumer genetics.
8. *Epigenomics*: It is the branch of genomics that deals with the study of supporting structure of genomes, including protein and RNA binders, alternative DNA structures, and modification on DNA.
9. *Nucleomics*: It is the branch of genomics that deals with the study of the complete set of genomic components that form the cell nucleus as a complex, dynamic biological system.

5.1.2 DNA sequencing

DNA sequencing is a process of identification of exact order of bases in a DNA strand. Once one of the bases in a pair is identified, there is no need to determine the order of other bases because the bases exist in a pair. It means that the determination of location and order of a base also determines the order of other bases (Metzker, 2005; Shendure & Ji, 2008).

Several methods of sequencing are used worldwide today and among them, the most popular method is sequencing by synthesis. In this method, DNA polymerase is used to synthesize a new strand of DNA of interest. Here the DNA polymerase incorporates with the individual nucleotides of new DNA strand which have been chemically tagged with a fluorescent level. When the reaction occurs, the nucleotide is excited by a light source and immediately fluorescent signal is emitted and detected. The fluorescent signal is different depending upon which of the four nucleotides was incorporated. This method can generate reads of 125 nucleotides in a row and billions of reads at a time (<http://www.genome.gov>). To understand and assemble the sequence of a gene (large piece of DNA), researchers read the sequence of overlapping segments which allows longer sequence to be assembled from shorter sequence (Table 5.1). In this process, each base needs to read several times in the overlapping segments to ensure accuracy (Shendure & Ji, 2008; Shendure, Balasubramanian, & Church, 2017).

DNA sequencing is used to search and identify any changes in genetic variations or mutations which might play role in the development of a disease in an organism. Any addition, deletion, or substitution in the base pairs might be harmful and could lead to serious diseases like cancer.

5.1.3 Research areas

Genomics has wide research areas, including structural and functional genomics, epigenomics, and metagenomics (Fig. 5.2).

5.1.3.1 Structural genomics

Structural genomics reports the three-dimensional structure of every protein encoded by a given genome (Brenner & Levitt, 2000; Marsden, Lewis, & Orengo, 2007). This genome-based approach help to fix the structure with a high-throughput method by a combination of both experimental and modeling approaches. Structural genomics tries to determine the structure of every protein encoded by a genome whereas traditional structural prediction focuses on a particular protein. With the availability of full genome sequencing, structural prediction can be done more quickly along with an addition of experimental and modeling approaches. It is because the accessibility of a large number of sequenced genomes and earlier solved protein structure assist scientists to model protein structures on the structure of earlier solved homologs. Structural genomics applies a large number of techniques to deal with the structure of a new protein including chemical and physical principles.

5.1.3.2 Functional genomics

Functional genomics deals with the function of DNA at the gene levels. It is a molecular biology that focuses on the dynamic aspects such as gene transcription, translation, and protein–protein interactions. Functional genomics uses vast wealth of data produced by genomic projects to describe gene and its interactions. The main characteristics include involving high-throughput methods rather than a more traditional gene-by-gene approach. The full knowledge of genomics creates the

TABLE 5.1 Chronology of sequencing events.

SN	Events	Date	References
1	Rosalind Franklin confirms the helical structure of DNA		Ankeny (2003)
2	James D. Watson and Francis Crick published structure of DNA	1953	Ankeny (2003)
3	Fred Sanger published amino acid sequence of insulin	1955	Ankeny (2003)
4	Robert W. Holley published the first nucleic acid sequence (the ribonucleotide sequence of alanine transfer RNA)	1964	Holley et al. (1965)
5	Marshall Nirenberg and Philip Leder determined the triplet nature of genetic code	1965	Nirenberg et al. (1965)
6	Walter Fiers was first to determine the sequence of a gene, the gene of bacteriophage Ms2 coat protein	1972	Min Jou, Haegeman, Ysebaert, and Fiers (1972)
7	Walter Fiers and his colleagues determined the complete nucleotide sequence of bacteriophage Ms2 RNA and simian virus 40	1976 and 1978	Fiers et al. (1976); Fiers et al. (1978)

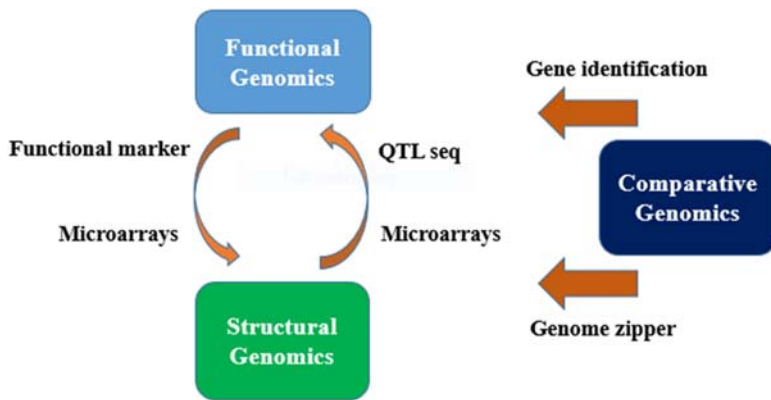


FIGURE 5.2 Interrelation among functional, structural, and comparative genomic approaches.

possibility for the field of functional genomics and mainly concerns with the pattern of gene expressions during various conditions. The important tools are microarrays and bioinformatics (Schuler & Reichhart, 2003).

5.1.3.3 Epigenomics

Epigenomics is the complete study of the epigenome. Epigenome is the complete set of modifications on the genetic material of a cell (Francis, 2011). Epigenetic modifications are reversible where the genes are expressed without altering the DNA sequence. It occurs on the cell's DNA or histones. The most important characteristics of epigenomics are DNA methylation and histone modification. Epigenetic modification plays important role in the gene expression and regulation and involves numerous cellular processes such as in differentiation and tumorigenesis. The study of epigenesis is being possible due to adaptation of genomic high-throughput assays (Callinan & Feinberg, 2006).

5.1.4 Model systems for the study of genome

5.1.4.1 Viruses and bacteriophages

Bacteriophages are playing a major role in the advancement of bacterial genetics and molecular biology. Earlier, bacteriophage was used to define gene structure and gene regulation. The first genome to be sequenced was also a bacteriophage. But bacteriophage did not lead genomic revolution and very recently the study on bacteriophage becomes prominent which enables to understand the mechanism underlying phase evolution. Bacteriophage genome sequence can be obtained through direct sequencing of isolated bacteriophages and can also be derived as a part of microbial genomes. Bacterial genome analysis shows that microbial DNA consists of prophage sequence and prophage-like elements (Canchaya, Proux, Fournous, Bruttin, & Brüßow, 2003; Metzker, 2005). The bacteriophage sequence mining helps to understand the role of prophage in shaping the bacterial genome. Therefore this method verified many bacteriophage groups and made a useful tool for predicting the relationships of prophage from bacterial genome (Fouts, 2006; McGrath & van Sinderen, 2007).

5.1.4.2 Cyanobacteria

Right now there are 24 cyanobacteria for which the total genome is sequenced, and 15 of which come from marine environment. These sequences could be used to infer important ecological and physiological characteristics of marine cyanobacteria. The increase in genome information could also be used to address global problems through a comparative approach. For example, identification of regulatory RNA genes enlight on evolutionary origin of photosynthesis, and estimation of contribution of horizontal gene transfer to the genome that has been analyzed (Herrero & Flores, 2008).

5.2 Development of genomic resources

5.2.1 Molecular markers

Molecular markers give highly precise results and are widely used in genomic research because of its wider acceptance. Molecular markers do not fluctuate with the environment, show high polymorphism, and most of them are codominant

in nature and give unbiased results. The first molecular marker discovered was hybridization-based DNA markers, that is, restriction fragment length polymorphism (RFLP) followed by PCR-based (polymerase chain reaction) markers, random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), and simple sequence repeat (SSR) marker for genotyping (Kumar, Rajendran, Kumar, Hamwiah, & Baul, 2015). Among various PCR-based markers, SSR markers significantly contributed to the development of different crops genome maps. Nowadays, these PCR-based markers are rapidly replaced by DNA chip-based markers like single-nucleotide polymorphisms (SNPs). SNPs are abundant in nature and are common even in legume genome (Chagne et al., 2007). Various technologies are available for evaluation of SNPs loci and many of these are amenable to automation for allele calling and data collection. The availability of sequence database has started to exploit them as an HTP (highthrough phenotyping) marker system for genome mapping studies (Kumar et al., 2015; Kumar, Gupta, Mishra, Modi, & Pandey, 2009). Recent attempts of resequencing allele to discover SNPs in crop like lentil have facilitated automated high-throughput genotyping platforms and therefore SNPs are emerged as latent markers for NGS (next generation sequencing) approaches (Kumar et al., 2015).

5.2.2 Transcriptome assemblies

Transcriptome assemblies provide good opportunity to identify expressed sequenced tags (ESTs) derived from SSR and SNP markers and intron-targeted primers (Table 5.2). Earlier Sanger method of dideoxynucleotide chain termination was used to sequence cDNA libraries and generate ESTs across several crops. ESTs are 150–400 bp short DNA sequences from a cDNA clone that corresponds to a particular mRNA. Recently, the development of HTP functional genomics approaches such as serial analysis of gene expression has led to the generation of more ESTs. The cDNA clone that corresponds to the ESTs of interest can be used as RFLP- or CAP-based (catabolite activator protein) markers (Varshney, Graner, & Sorrells, 2005). Thus EST sequence data is also used for identification of SSRs and SNPs. Earlier when ESTs sequence was not available, the development of SSR and SNP markers was expensive, tedious and required high-resource laboratories, but nowadays any individual can download the database and can use some special bioinformatics program like MISA (a web server for microsatellite prediction) for SSR detection and Snipper for SNP discovery (Varshney et al., 2005; Thiel, Michalek, Varshney, & Graner, 2003).

5.2.3 Biparental mapping populations

Mapping populations are developed to identify key traits in a crop. Mapping populations are of different types like recombinant inbred line (RIL), near-isogenic line (NIL), multiparent advanced generation intercross (MAGIC), nested association mapping (NAM), and F2 populations. Identification of markers linked to the genes/QTL (quantitative trait

TABLE 5.2 Transcription factor database in plants (Mochida & Shinozaki, 2010).

Database	Species	URL
RARTF	<i>Arabidopsis</i>	http://rarge.gsc.riken.jp/rartf/
AGRIS, AtTFDB	<i>Arabidopsis</i>	http://arabidopsis.med.ohio-state.edu/AtTFDB/
DATF	<i>Arabidopsis</i>	http://datf.cbi.pku.edu.cn/
DRTF	Rice	http://drtf.cbi.pku.edu.cn/
DPTF	Poplar	http://dptf.cbi.pku.edu.cn/
TOBFAC	Tobacco	http://compsysbio.achs.virginia.edu/tobfac/
SoybeanTFDB	Soybean	http://soybeantfdb.psc.riken.jp/
PlantTFDB	22 plant species	http://plantfdb.cbi.pku.edu.cn/
PlnTFDB	20 plant species	http://plntfdb.bio.uni-potsdam.de/v3.0/
GRASSIUS, GrassTFDB	Maize, rice, sorghum, sugarcane	http://grassius.org/grasstfdb.html
LegumeTFDB	Soybean, <i>Lotus japonicus</i> , <i>Medicago truncatula</i>	http://legumetfdb.psc.riken.jp/
DBD	> 700 species	http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home

loci) governing target traits help in the development of genotype having high biomass at early stage. With the rapid generation advancement technology and speed breeding, four to five generations of annual crops per year can be grown which help in the development of much-needed genomic resources for genomics-enabled improvement (Kumar et al., 2015; Watson, Ghosh, & Williams, 2018). Biparental populations have three major advantages (Scott, Ladejobi, & Amer, 2020):

1. Relative simplicity of construction, that is, only two generations are needed for F2 populations and only six generations of inbreeding in self-pollinated species whose genomes are fixed are needed for making RILs.
2. The probability of QTL detection becomes high because of all allele frequencies which are typically close to the optimal value of 50%.
3. The rate of linkage disequilibrium decay within the chromosomes becomes low because only one or two recombinants per chromosomes arm mean only a few hundred genotyped markers are needed to map QTL.

5.2.4 Genetic linkage maps

Genetic mapping (linkage mapping) was first begun by Zamir and Ladizinsky (1984) and the first map comprising DNA-based markers was given by Havey and Muehlbauer (1989). After that, several maps were produced. The increase in number of markers was high across many crops, including lentil with the discovery of PCR-based markers (Kumar et al., 2014). Mapping populations could be inter- or intraspecific but intraspecific mapping populations have been found more practical utility in QTL identification than to tag desirable genes of interest (Kumar et al., 2015).

5.2.5 Comparative genome mapping

Different levels of genome conservation among crop species have been demonstrated by comparative genome mapping during the course of evolution (Zhu, Choi, Cook, & Shoemaker, 2005). PCR-based markers have improved the transferability genetic information among species through comparative genomics and have facilitated the start-up of phylogenetic relationships in plant species. Comparative genomics provides opportunity for study of genetic diversity (Hardison, 2003; Kumar et al., 2015).

5.2.6 Functional genomics

Genomic maps are used to identify genes/QTL of interest. Similarly, gene cloning approach helps to characterize and reveal functions of gene/QTL being identified (Table 5.3). Thus the accumulative knowledge of gene cloned in a crop facilitates the development of functional markers for MAS (marker assisted selection). With functional genomic approaches, genes expressing differentially in contrasting genotypes can also be identified. Microarray is also used to identify gene network underlying the expression of important plant traits (Kumar et al., 2015; Schuler & Reichhart, 2003).

5.3 Application of genomic resources for crop improvement

5.3.1 Genetic fingerprinting

Several crops' genetic diversity has been studied using various molecular markers like RFLP, AFLP, and RAPD to access genetic diversity and phylogenetic analysis within and among crop species. The diversity analysis and gene mapping help in genetic characterization of genotypes. Cluster analysis could be used to group the accessible germplasm into certain clusters (Kumar et al., 2015).

5.3.2 Hybrid testing

When plants and its flowers are of small in size, then crossing becomes difficult and it increases the chance of selfing. Similarly, differentiation of F1 plants from selfed ones also becomes difficult due to low phenotypic diversity between the parents (Kumar et al., 2015). Molecular markers can reduce the time and money required to grow a population from selfed or admixed plants and increase the efficiency of plant breeders in the selection of recombinant plants.

TABLE 5.3 Crop of specific traits along with its particular gene/locus (Tao, Zhao, Mace, Henry, & Jordan, 2019).

Species	Gene (Locus)	Trait
Rice	GW5/qSW5/GSE5	Grain size
Rice	GL7	Grain size
Rice	SGDP7	Grain size, grain number, yield
Rice	Pikm1-TS	Blast resistance
Rice	Pikm2-TS	Blast resistance
Rice	Sub1A	Submergence tolerance
Rice	Pup1	Phosphorus-starvation tolerance
Rice	SNORKEL1	Deepwater response
Rice	SNORKEL2	Deepwater response
Rice	qPE9-1	Plant architecture
Rice	Pi21	Blast disease
Rice	Sc	Hybrid male sterility
Rice	DPL1/DPL2	Hybrid male sterility
Rice	S27/S28	Hybrid male sterility
Rice	OsSh1	Shattering
Maize	KRN4	Kernel row number
Maize	ZmCCT10	Photoperiod sensitivity
Maize	TB1	Apical dominance
Maize	Vgt1	Flowering time
Maize	qHSR1	Resistance to head smut
Maize	Scmv1	Resistance to sugarcane mosaic virus
Maize	ZmCCT9	Photoperiod sensitivity
Maize	MATE1	Aluminum tolerance
Sorghum	LGS1	Resistance to Striga
Sorghum	Sh1	Shattering
Sorghum	SbMATE	Aluminum tolerance
Wheat	Lr10	Leaf rust
Wheat	Yr36	Stripe rust
Wheat	Tsn1	Tan spot and Stagonospora nodorum blotch
Wheat	FR-2	Cold tolerance
Wheat	Vrn-A1	Vernalization
Wheat	Ppd-B1	Photoperiod sensitivity
Wheat	Rht-D1b	Plant height
Barley	FR-H2	Frost resistance
Barley	Bot1	Boron-toxicity tolerance
Barley	HvFT1	Flowering time
Soybean	GmCHX1	Salt tolerance
Soybean	Rhg1	Resistance to cyst nematode

(Continued)

TABLE 5.3 (Continued)

Species	Gene (Locus)	Trait
Cucumber	F	Sexual production
Cucumber	Tu	Tuberculate fruit
Opium poppy	NA	Production of noscapine
Tomato	SUN	Elongated fruit shape
Tomato	TMF	Flowering
Potato	R1	Resistance against late blight
Potato	ELR	Resistance against late blight

5.3.3 Marker-assisted selection

Marker-assisted selection uses molecular markers linked with desirable gene/QTL in many crops. Different types of mapping populations are used for the identification of QTL. F2 population is heavily used for the identification of major traits in lentil but it only identifies major QTL. Because QTL is highly influenced by both genotype and the environment but RIL or NIL populations are more suitable to identify quantitative traits and dissect their components. The flanking markers have also been found promising for MAS and pyramiding of potentially different resistance genes into elite background which are resistant throughout the cropping season (Kumar et al., 2009; Kumar et al., 2015).

5.3.4 Gene trait association analysis using natural diverse populations

Biparental mating approach could cause high chances of segregation distortion by favoring one parental allele over the other. The molecular markers that show polymorphism within the interspecific populations might not show polymorphism at the species level as genetic background affects their utility in MAS process. An alternative approach like association mapping could address the shortcomings of biparental mapping. Association mapping does marker–trait association and identifies QTL with high resolution using historical recombination in natural populations, landraces, breeding materials, and varieties. Association mapping is of two types, genome-wide association studies (GWAS) and candidate gene association mapping (Kumar et al., 2015). Any crop used for association mapping must be rich with genomic resources.

5.3.5 Genetic transformations

Genetic transformation is a biotechnological approach that transfers functional genes to the target species that are not available in the crossable gene pool. Thus for desired genetic manipulation, cloned genes are important genetic resources. Mainly two approaches particle bombardment and *Agrobacterium tumefaciens* infection methods are used to introduce genes with novel functions. With the availability of sequence information obtained through the database, transformation systems become very useful to study gene function via RNA interference (knockout), T-DNA insertion, or transforming a genotype which lack particular gene. Thus a robust transformation system combined with a protocol to regenerate complete fertile plant from transformed cell is essential to fully study plant gene functions (Burt, 2003; Kumar et al., 2015).

5.4 Genome analysis

Genome analysis of an organism involves major three components, that is, DNA sequencing, assembling (assembling of DNA sequence to create representation of original chromosome), and annotation and analysis of that representation (Table 5.4).

TABLE 5.4 Integrative databases in plants (Mochida & Shinozaki, 2010).

Database name	Species	URL
TAIR	<i>Arabidopsis</i>	http://www.arabidopsis.org/
SIGnAL	<i>Arabidopsis</i>	http://signal.salk.edu/
RARGE	<i>Arabidopsis</i>	http://rarge.psc.riken.jp/
Rice Genome Annotation Project	Rice	http://rice.plantbiology.msu.edu/
RAP-DB	Rice	http://rapdb.dna.affrc.go.jp/
SOL Genomics Network	Solanaceae	http://solgenomics.net/
Gramene	Gramineae	http://www.gramene.org/
GrainGenes	Triticeae and Avena	http://wheat.pw.usda.gov/GG2/index.shtml
SoyBase	Soybean	http://www.soybase.org/
MaizeGDB	Maize	http://www.maizegdb.org/
CyanoBase	<i>Cyanobacteria</i>	http://genome.kazusa.or.jp/cyanobase/
GDR (Genome Database for Rosaceae)	Rosaceae	http://www.bioinfo.wsu.edu/gdr/
Brassica Genome Gateway	<i>Brassica</i>	http://brassica.bbsrc.ac.uk/
Cucurbit Genomics Database	Cucurbitaceae	http://www.icugi.org/
Phytozome	Plant species (whole-genome data available)	http://www.phytozome.net/
PlantGDB	Plant species (whole-genome and/or large-scale EST data available)	http://www.plantgdb.org/
Ensembl Plants	Plant species (whole genome data available)	http://plants.ensembl.org/index.html
ChloroplastDB	Plant species (Chloroplast genome data available)	http://chloroplast.cbio.psu.edu/
KEGG Plant	Plant species (whole-genome and/or large-scale EST data available)	http://www.genome.jp/kegg/plant/

EST, Expressed sequenced tags.

5.4.1 Sequencing

DNA sequencing approaches fall into two broad categories: shotgun and high-throughput (next-generation) sequencing (Shendure & Ji, 2008; Shendure et al., 2017; Shi & Anderson, 2003).

5.4.1.1 Shotgun sequencing

Shotgun sequencing is a sequencing method designed for the analysis of DNA sequences longer than 1000 base pairs and might include sequencing of entire chromosomes (Staden, 1979). Because gel electrophoresis sequencing is used to sequence only short base pairs (100–1000 base pairs), longer DNA sequences are broken into small random fragments and are then sequenced to get new reads. Several rounds of fragmentation and sequencing are done to confirm reads of multiple overlapping DNA. After that, computer programs use overlapping ends of different reads that assemble them in a continuous sequence (Venter, Adams, & Sutton, 1998).

Shotgun sequencing is a random sampling process that requires large amount of sampling to ensure that a given nucleotide is represented in the reconstructed sequence. Shotgun sequencing is a classical chain termination method (Sanger method) that is based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication (Sanger & Coulson, 1975). But nowadays, this method is incorporated with high-throughput sequencing for large-scale automated genome analysis. Still Sanger method is widely used worldwide for small-scale genome projects to obtain contiguous DNA sequence. Chain termination method requires a single-stranded

DNA template, DNA primer, DNA polymerase, normal deoxynucleoside triphosphates (dNTPs), and modified nucleotides (dideoxynucleotides) which terminate DNA strand elongation. These chain-terminating nucleotides lack a 3-OH group which is required for the formation of a phosphodiester bond between two nucleotides and cause DNA polymerase to cease extension of DNA when a ddNTP is incorporated. The ddNTPs might be radioactively or fluorescently labeled for the detection of DNA sequencers (Pevsner, 2009). These machines can sequence up to 96 DNA samples in a single batch in up to 48 runs a day.

5.4.1.2 High-throughput sequencing

High-throughput sequencing is a low-cost sequencing technology that parallelizes the sequencing process and produces millions of sequence at once (Church, 2006; Hall, 2007). This sequencing technique reduces the cost of DNA sequencing and produces all the possible results with standard dye terminator methods. Nearly 5,00,000 sequencing by synthesis operations might be run in parallel by ultrahigh-throughput sequencing (ten Bosch & Grody, 2008; Tucker, Marra, & Friedman, 2009).

High-throughput sequencing can be done in two ways: illumina dye sequencing and ion semiconductor sequencing. The illumina dye sequencing method is developed by Pascal Mayer and Laurent Farinelli in 1996 at the Geneva Biomedical Research Institute. This method is based on reversible dye terminators where DNA molecules and primers are attached on a slide and are amplified with polymerase which forms local clonal bodies or DNA colonies. To determine the sequence of DNA colonies, four types of reversible terminator bases are added and non-incorporated nucleotides are washed away. Then the DNA chains are extended with one nucleotide at a time and image acquisition is performed later which allows large arrays of DNA colonies captured by sequential images taken from a single camera. The optimal throughput and unlimited sequencing capacity with an optimal configuration are possible only by decoupling the enzymatic reaction and image capture capacity. The optimal throughput of the instrument depends on A/D (addition/deletion) conversion rate of the camera where the camera takes image of the fluorescently labeled nucleotide after that, the dye along with the terminal 3' blocker is chemically removed from the DNA which allows the next cycle (Anders, Theodor Pyl, & Huber, 2015; Mardis, 2008).

The ion semiconductor sequencing is based on standard DNA replication which measures the release of hydrogen ion each time a base is incorporated. A template DNA is flooded with a single nucleotide in a micro-well where a hydrogen ion will be released if the nucleotide is complementary to the template strand. This release of hydrogen ion triggers an ISFET (ion-sensitive field-effect transistor) ion sensor. Multiple nucleotides will be incorporated in a single flood cycle if a homopolymer is present in the template sequence and the detected electrical signal will be higher.

5.4.2 Assembly

Sequence assembly is aligning and merging fragments of longer DNA sequence to reconstruct original sequence (Pevsner, 2009). The current DNA sequencing technology reads small piece of sequences between 20 and 1000 bases and cannot read much longer sequence (whole genome) as a continuous sequence. But the third-generation sequencing technology like PacBio routinely generates sequencing reads >10 kb in length but they are having high error (approx. 15%) (<https://www.pacb.com/>). The short fragments of nucleotide called reads result from shotgun DNA sequencing.

5.4.2.1 Assembly approaches

Assembly approaches can be categorized into two types: de novo and comparative assembly. De novo assembly is used for the genomes that are not similar to any sequenced in the past whereas comparative assembly uses the existing sequence of a closely related organism as a reference (Pop, 2009). De novo assembly is computationally difficult and is less favorable for short-read NGS technologies (Rahimi-Vahed, Rabbani, Tavakkoli-Moghaddam, Torabi, & Jolai, 2007).

5.4.2.2 Finishing

The genomes have single contiguous sequence with no ambiguities representing each replicon.

5.4.3 Annotation

The genome sequence assembly needs additional analysis to get high value (Pevsner, 2009). Therefore genome annotation is needed. Genome annotation is the process of attaching biological information to the DNA sequences which consists of three steps (Stein, 2001):

1. Identification of portion of the genome which do not code for proteins
2. Identification of elements on the genome through a process called gene prediction and
3. Attachment of biological information to these elements

Nowadays, automatic annotation tools are available which perform above mentioned steps in silico. Earlier, manual annotation (curation) applied that involved human expertise and experimental verification (Brent, 2008). Both these approaches coexist and complement each other in the same annotation pipeline. Earlier, basic level of annotations was done using BLAST for finding similarities and then annotating genomes based on homologs (Pevsner, 2009). But recently, more information is added to the annotation platform. These more information allow manual annotators to deconvolute discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integration of some other resources to provide genome annotations through their subsystems approaches. Similarly other databases rely on curated data sources and a range of software tools in their automated genome annotation pipelines (Flicek et al., 2013). In structural annotation, identification of genomic elements, primarily ORFs (open reading frames) and their localization or gene structure, are done whereas in functional annotation biological information are attached to the genomic elements.

5.5 Applications of genomics

Genomics has wide applications in several fields, including medicine, biotechnology, anthropology, and social sciences (Barnes & Dupré, 2008). The major fields and its applications are discussed in the following sections.

5.5.1 Genomics in medicine

Clinicians and biomedical researchers are able to increase the amount of genomic data drastically collected on the large study populations through next-generation genomic technologies (Hudson, 2011). These huge amounts of genomic data are combined with new informatics approaches which integrate many kinds of data in disease research, and it allows the researcher to better understand the genetic bases of drug response and disease (O'Donnell & Nabel, 2011; Lu, Goldstein, Angrist, & Cavalleri, 2014). Euan Ashley applied genome to medicine and developed the first tool for the medical interpretation of a human genome (Ashley et al., 2010; Dewey et al., 2011; Dewey et al., 2014).

5.5.2 Genomics in synthetic biology and bioengineering

The vast knowledge of genomics helps in sophisticated applications of synthetic biology (Church & Regis, 2012). The creation of partially synthetic species of bacterium *Mycoplasma laboratorium*, derived from the genome of *Mycoplasma genitalium*, was possible at J Craig Venter Institute in 2010 (Baker, 2011).

5.5.3 Conservation genomics

The information gathered through genomic sequencing is used to better evaluate the genetic factors key to species conservation. For example, genetic diversity of a population is used to understand whether an individual is heterozygous for a recessive inherited genetic disorder (Frankham, 2010). Genomic data are also used to evaluate the effects of evolutionary processes and to detect pattern in variation throughout a given population.

5.6 Next-generation genomics for crop improvement

Resequencing of genome-wide sequence variation significantly improves the availability of information that can be used to develop genetic markers and, therefore, proceed the genetic mapping of agronomic traits. For example, only less than 500 SNP markers were available in 2008, and SNP markers increased to 1536 in 2010, 10,000 in 2011, and more than 90,000 in 2012 (Allen et al., 2011; Chao et al., 2008). This high-density SNP information is proving highly useful to different systems like QTL mapping in biparental crosses and RILs, GWASs, mapping QTL in MAGIC lines.

TABLE 5.5 Progress in crop genome sequencing (Bevan & Uauy, 2013).

Species (common name)	Genome size	Ploidy	Sequence strategy	Publication date	Assembly features
<i>Oryza sativa</i> (rice)	389 Mb	$2n = 2x = 24$	BAC physical map, Sanger sequencing	Aug 2005	Essentially complete chromosome arm coverage
<i>Populus trichocarpa</i> (black cottonwood)	550 Mb	$2n = 2x = 8$	BAC physical map, WGS, Sanger sequencing	Sep 2006	2447 c scaffolds containing 410 Mb, 82% of sequence genetically anchored
<i>Vitis vinifera</i> (pinot noir grape)	475 Mb	$2n = 2x = 36$	WGS, Sanger sequencing	Sep 2007	3514 c supercontigs containing 487 Mb, 69% of sequence genetically anchored
<i>Sorghum bicolor</i> (sorghum)	700 Mb	$2n = 2x = 20$	WGS, Sanger sequencing	Jan 2009	229 scaffolds containing 97% of the genome, 88% of sequence genetically anchored
<i>Zea mays</i> (maize)	2300 Mb	$2n = 2x = 20$, one a WGD allotetraploid	BAC physical map, BAC sequence 4–6x deep	Nov 2009	2048 Mb in 125,325 b contigs forming 61,161 scaffolds
<i>Glycine max</i> (soybean)	1115 Mb	Two WGD $2n = 2x = 40$ allopolyploid	WGS, Sanger sequencing	Jan 2010	397 scaffolds containing 85% of the genome, 98% of sequence genetically anchored
<i>Malus x domestica</i> (apple)	750 Mb	One WGD $2n = 2x = 34$	WGS, Sanger, Roche 454	Oct 2010	1629 c metacontigs containing 80% of the genome, 71% of sequence genetically anchored
<i>Theobroma cacao</i> (cacao)	430 Mb	$2n = 2x = 20$	WGS, Sanger, Illumina, Roche 454	Dec 2010	524 scaffolds containing 80% of the genome, 67% of sequence genetically anchored
<i>Fragaria vesca</i> (woodland strawberry)	240 Mb	$2n = 2x = 14$	WGS, Roche 454, Illumina, SOLiD	Dec 2010	272 scaffolds containing 95% of the genome, 94% of sequence genetically anchored
<i>Phoenix dactylifera</i> (date palm)	658 Mb	$2n = 2x = 36$	WGS, Illumina	June 2011	57,277 scaffolds containing 60% of the genome
<i>Solanum tuberosum</i> (potato)	844 Mb	$2n = 4x = 48$	Double monoploid DM and diploid RH, WGS, Illumina, Roche 454	July 2011	443 superscaffolds containing 78% of the genome, 86% of the assembly genetically anchored
<i>Brassica rapa</i> (Chinese cabbage)	485 Mb	Three WGD $2n = 2x = 20$	WGS, Illumina, BAC end Sanger sequencing	Aug 2011	288 Mb in scaffolds, 90% of the assembly genetically anchored
<i>Medicago truncatula</i> (alfalfa relative)	375 Mb	WGD $2n = 2x = 16$	BAC physical map, Sanger, Illumina	Dec 2011	8 pseudomolecules containing 70% of the genome, 100% in optical map
<i>Manihot esculenta</i> (cassava)	770 Mb	$2n = 2x = 36$	WGS, Roche 454, BAC end Sanger sequencing	Jan 2012	12,977 scaffolds containing 80% of the genome
<i>Cajanus cajan</i> (pigeonpea)	833 Mb	$2n = 2x = 22$	WGS, Illumina	Jan 2012	137,542 scaffolds containing 73% of the genome

(Continued)

TABLE 5.5 (Continued)

Species (common name)	Genome size	Ploidy	Sequence strategy	Publication date	Assembly features
<i>Setaria italica</i> (foxtail millet)	500 Mb	$2n = 2x = 18$	WGS, Sanger, Illumina, BAC end sequence	May 2012	597 scaffolds containing 80% of the genome, 99% of the assembly genetically anchored
<i>Solanum lycopersicum</i> (tomato)	900 Mb	$2n = 2x = 24$	WGS, Roche 454, Illumina and SOLiD, BAC end Sanger sequencing	May 2012	91 scaffolds containing 85% of the genome, 99% of the assembly genetically anchored
<i>Cucumis melo</i> (melon)	312 Mb	Three WGD $2n = 2x = 24$	WGS, Roche 454, BAC end sequencing	July 2012	1584 scaffolds containing 83% of the genome, 88% of the assembly genetically anchored
<i>Musa acuminata</i> (Cavendish banana)	523 Mb	$2n = 2x = 22$	WGS, Roche 454, Sanger, Illumina	Aug 2012	24,425 contigs containing 90% of the genome, 70% of the assembly genetically anchored
<i>Citrus sinensis</i> (Valencia sweet orange)	367 Mb	$2n = 2x = 18$	Dihaploid WGS, Illumina	Jan 2013	4,811 scaffolds containing 82% of the genome, 73% of the assembly genetically anchored
<i>Gossypium raimondii</i> (D genome cotton)	880 Mb	$2n = 2x = 26$	WGS, Illumina	Aug 2012	4,715 scaffolds containing 85% of the genome, 73% of the assembly genetically anchored
<i>Hordeum vulgare</i> (barley)	5100 Mb	$2n = 2x = 14$	WGS, Illumina, BAC physical map, BAC sequence (Roche 454, Illumina)	Nov 2012	Physical map (4.98 Gb), BAC sequence (1.13 Gb), WGS assemblies (1.9 Gb); integrated by physical map and syntenic order
<i>Triticum aestivum</i> (bread wheat)	17,000 Mb	$2n = 6x = 42$ allopolyploid	WGS, Roche 454	Nov 2012	Orthologous group assembly, 437 Mb
<i>G. raimondii</i> (D genome cotton)	880 Mb	$2n = 2x = 26$	WGS, Sanger, Roche 454, Illumina	Dec 2012	1084 scaffolds containing 86% of the genome, 98% anchored and oriented to genetic map
<i>Gossypium hirsutum</i> (upland cotton)		AtDt allopolyploid	Illumina		82x coverage
<i>Cicer arietinum</i> (chickpea)	738 Mb	$2n = 2x = 16$	WGS, Illumina BAC end sequence	Jan 2013	7163 scaffolds containing 64% of the genome
<i>Phyllostachys heterocycla</i> (bamboo)	2 Gb	$2n = 2x = 48$	WGS, Illumina BAC end sequence	Apr 2013	80% of the 2.05 Gb assembly maps to 5499 scaffolds of less than 62 kb
<i>Picea abies</i> (Norway spruce)	20,000 Mb	$2n = 2x = 24$	Fosmid pools with both haploid (megagametophyte) and diploid WGS	May 2013	Merged assembly 12.0 Gb, with 4.3 Gb in ≥ 10 kb scaffolds
<i>Pinus taeda</i> (Loblolly pine)	24,000 Mb	$2n = 2x = 24$	WGS single haploid megagametophyte assembly	In progress	

(Continued)

TABLE 5.5 (Continued)

Species (common name)	Genome size	Ploidy	Sequence strategy	Publication date	Assembly features
<i>Miscanthus</i> sp. (elephant grass)	1500 Mb	One WGD, diploid progenitors $2n = 2x = 38$	WGS	In progress	
<i>Elaeis guineensis</i> , <i>Elaeis oleifera</i> (oil palm)	1890 Mb	$2n = 2x = 32$ commercial F1 hybrids	WGS, BAC physical maps	In progress	
<i>Saccharum officinarum</i> x <i>S. spontaneum</i> (sugarcane)	>15,000 Mb	Diploid progenitors $x = 10$; $2n = 80$; $x = 8$; $2n = 40-128$	WGS	In progress	

a WGD allopolyploids have a whole-genome duplication in recent lineage. b A contig is an unambiguous linear assembly of sequences with no physical gaps in coverage, but which can contain errors. c The terms supercontig, scaffold or metacontig are used interchangeably to describe a set of contigs that are linked by a known physical distance but that contain sequence gaps. These scaffolds are usually created using mate-pair reads and BAC end sequences. d Pseudomolecule is a term applied to a chromosome-scale assembly of contigs and scaffolds that is anchored to a long-range framework using genetic markers and other chromosome features, including cytogenetic features and deletions. DM, disease causing Mutations; RH, rhesus factor; BAC, bacterial artificial chromosome; WGS, whole genome sequencing.

The resequencing approaches identify loci and causal genes for traits with relatively large phenotypic effects. By then the genomic segments which contain desired allelic variation can be bred and combined in a single genetic background by using markers to track the segments through marker-assisted breeding.

Some agronomical important traits like yield are the cumulative small effects of several loci which cannot be identified through QTL or GWAS approaches and their pyramiding through MAS will also be ineffective. Thus breeder addressed these problems by developing the knowledge base of the associations of polymorphic markers with phenotypes in breeding populations (Bevan & Uauy, 2013). These associations are used to develop a breeding model in which the frequency of desired marker allele is optimized and, therefore, maximize the estimated breeding value (Xu, 2003; Heffner, Sorrells, & Jannink, 2009). Therefore the rate of selection cycle is multiplied to accumulate favorable alleles that are associated with desired phenotypes though no relationship between particular gene and the phenotype is established. This approach is termed genomic selection (Eathington, Crosbie, Edwards, Reiter, & Bull, 2007). Genomic selection is influenced by next-generation sequencing of parental lines in many ways, by continuing to identify polymorphism throughout the genome in both genic and intergenic regions which remove any limitations on marker density, by providing estimates of gene expression levels, and by providing information on the epigenetic states of genes which are genetic features and have predictive power for complex traits (Bevan & Uauy, 2013).

Conventional breeding uses naturally available allelic variation for crop improvement. But sequence variation can be created artificially by using ethyl methane sulfonate to alkylate bases. By then, TILLING (targeted induced local lesions in genomes) is used to screen the changes of bases in the genes of interest to access gene function and to apply this for allele breeding (Uauy et al., 2009). Now it is possible to use genome capture to sequence an entire mutant population and even the complex polyploidy genomes like wheat.

Genetic modifications or transfer of genes from one organism to the other by *A. tumefaciens* is fully developed technology and is well adapted for use in many of the crop species (Table 5.5). The precise modification in gene sequence using zinc finger nucleases (ZNS) can be applied which recognize specific gene sequence with a target location in maize (Shukla et al., 2009). CRISPR (clustered regularly interspaced short palindromic repeats), a new type of precision tool for genetic engineering, is developed from prokaryotes, which is guided to specific target sequences for cleavage by an RNA molecule (Jinek et al., 2012; Mali et al., 2013). Similarly, several types of genome editing are now possible such as simultaneous editing of multiple sites, inducing deletions, and inserting new sequences by nick-mediated repair mechanism (Bevan & Uauy, 2013).

5.7 Genomic features for future breeding

The scope of genetics has radically altered with the introduction of genomics which provides a landscape of ordered genes and their epigenetic states, makes access to enormous range of genetic variation, and possesses the potential to measure gene expression with high precision directly. Genomics also facilitate systemic comparison of gene functions across sequenced genomes which gives abundant knowledge of gene functions and network obtained in experimental species and this can be used for crop improvement (Bustamante et al., 2011; Ogbe et al., 2016). The biological knowledge and models of network across species can be integrated through a suitable cyber infrastructure in a two-way flow from crop to experimental species and reverse again, which will generate new ways of knowledge which can be applied for crop improvement. One layer is given by ENCODE analysis which helps in interpretation of gene function and variation and provides new information for the prediction of phenotype from the genotype (The ENCODE Project Consortium, 2004). The other layer information is given by system-level integration of gene function into network, for example, controlling flowering time with response to day length. This network is identified in rice and *Arabidopsis*. Evolutionary processes like gene duplication and footprints of domestication can be mapped to network such as those controlling flowering time (Lander, 1996; Yan et al., 2006; Higgins, Bailey, & Laurie, 2010). These types of system breeding approaches use diverse genomic information to increase the precision by which phenotype is predicted from genotype and thereby speed up crop improvement (Bevan & Uauy, 2013).

References

- Allen, A. M., Barker, G. L. A., Berry, S. T., Coghill, J. A., Gwilliam, R., Kirby, S., ... Edwards, K. J. (2011). Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, 9, 1086–1099.
- Anders, S., Theodor Pyl, P., & Huber, W. (2015). HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2), 166–169. Available from <https://doi.org/10.1093/bioinformatics/btu638>.
- Ankeny, R. A. (2003). Sequencing the genome from nematode to human: changing methods, changing science. *Endeavour*, 27(2), 87–92. Available from [https://doi.org/10.1016/S0160-9327\(03\)00061-9](https://doi.org/10.1016/S0160-9327(03)00061-9).
- Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., ... Altman, R. B. (2010). Clinical assessment incorporating a personal genome. *The Lancet*, 375(9725), 1525–1535. Available from [https://doi.org/10.1016/S0140-6736\(10\)60452-7](https://doi.org/10.1016/S0140-6736(10)60452-7).
- Baker, M. (2011). Synthetic genomes: The next step for the synthetic genome. *Nature*, 473(7347), 405–408. Available from <https://doi.org/10.1038/473403a>.
- Barnes, B., & Dupré, J. (2008). *Genomes and what to make of them*. Chicago: University of Chicago Press, ISBN 978-0-226-17295-8.
- Bevan, M. W., & Uauy, C. (2013). Genomics reveals new landscapes for crop improvement. *Genome Biology*, 14, 206. Available from <http://genome-biology.com/2013/14/6/206>.
- Brenner, S. E., & Levitt, M. (2000). Expectations from structural genomics. *Protein Science*, 9(1), 197–200. Available from <https://doi.org/10.1110/ps.9.1.197>.
- Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews. Genetics*, 9(1), 62–73. Available from <https://doi.org/10.1038/nrg2220>.
- Burt, A. (2003). Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proceedings of the Royal Society B: Biological Sciences*, 270, 921–928. Available from <https://doi.org/10.1098/rspb.2002.2319>.
- Bustamante, C., De La Vega, F., & Burchard, E. (2011). Genomics for the world. *Nature*, 475, 163–165. Available from <https://doi.org/10.1038/475163a>.
- Canchaya, C., Proux, C., Fournous, G., Bruttin, A., & Brüssow, H. (2003). Prophase genomics. *Microbiology and Molecular Biology Reviews*, 67(2), 238–276. Available from <https://doi.org/10.1128/MMBR.67.2.238-276.2003>.
- Chagne, D., Carlisle, C. M., Blond, C., Volz, R. K., Whitworth, C. J., Oraguzie, N. C., et al. (2007). Mapping a candidate gene (MdMYB10) for red flesh and foliage colour in apple. *BMC Genomics*, 8, 212. Available from <https://doi.org/10.1186/1471-2164-8-212>.
- Chao, S., Zhang, W., Akhunov, E., Sherman, J., Ma, Y., Luo, M. C., & Dubcovsky, J. (2008). Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Molecular Breeding*, 23, 23–33.
- Church, G. M. (2006). Genomes for all. *Scientific American*, 294(1), 46–54. Available from <https://doi.org/10.1038/scientificamerican0106-46>.
- Church, G. M., & Regis, E. (2012). *Regenesis: how synthetic biology will reinvent nature and ourselves*. New York: Basic Books, ISBN 978-0-465-02175-8.
- Cooper, R. S., Kaufman, J. S., & Ward, R. (2003). Race and genomics. *The New England Journal of Medicine*, 348(12), 1166–1170.
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, 29, 229–237.
- Dewey, F. E., Chen, R., Cordero, S. P., Ormond, K. E., Caleshu, C., Karczewski, K. J., ... Ashley, E. A. (2011). Phased whole genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genetics*, 7(9), e1002280. Available from <https://doi.org/10.1371/journal.pgen.1002280>.

- Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., . . . Quertermous, T. (2014). Clinical interpretation and implications of whole-genome sequencing. *JAMA: the Journal of the American Medical Association*, *311*(10), 1035–1045. Available from <https://doi.org/10.1001/jama.2014.1717>.
- Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S., & Bull, J. K. (2007). Molecular markers in a commercial breeding program. *Crop Science*, *47*, 154–163.
- Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., . . . Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, *260*(5551), 500–507. Available from <https://doi.org/10.1038/260500a0>.
- Fiers, W., Contreras, R., Haegemann, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., . . . Ysebaert, M. (1978). Complete nucleotide sequence of SV40 DNA. *Nature*, *273*(5658), 113–120. Available from <https://doi.org/10.1038/273113a0>.
- Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., . . . Searle, S. M. J. (2013). Ensembl 2013. *Nucleic Acids Research*, *41*(Database issue): 48–55. Available from <https://doi.org/10.1093/nar/gks1236>.
- Fouts, D. E. (2006). Phage finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, *34*(20), 5839–5851. Available from <https://doi.org/10.1093/nar/gkl732>.
- Francis, R. C. (2011). *Epigenetics: the ultimate mystery of inheritance*. New York: WW Norton, ISBN 978-0-393-07005-7.
- Frankham, R. (2010). Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, *143*(9), 1922–1923. Available from <https://doi.org/10.1016/j.biocon.2010.05.011>.
- Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology*, *210*(9), 1518–1525. Available from <https://doi.org/10.1242/jeb.001370>.
- Hardison, R. C. (2003). Comparative genomics. *PLoS Biology*, *1*(2), e58. Available from <https://doi.org/10.1371/journal.pbio.0000058>.
- Havey, M. J., & Muehlbauer, F. J. (1989). Linkages between restriction fragment length, isozyme and morphological markers in lentil. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *77*, 395–401. Available from <https://doi.org/10.1007/bf00305835>.
- Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*, 1.
- The cyanobacteria: Molecular biology In A. Herrero, & E. Flores (Eds.), *Genomics and Evolution* (1st ed.). Caister Academic Press, ISBN 978-1-904455-15-8.
- Higgins, J. A., Bailey, P. C., & Laurie, D. A. (2010). Comparative genomics of flowering time pathways using *Brachypodium distachyon* as a model for the temperate grasses. *PLoS One*, *5*, e10065.
- Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., . . . Zamir, A. (1965). Structure of a ribonucleic acid. *Science (New York, N.Y.)*, *147*(3664), 1462. Available from <https://doi.org/10.1126/science.147.3664.1462>.
- Hudson, K. L. (2011). Genomics, health care, and society. *The New England Journal of Medicine*, *365*(11), 1033–1041. Available from <https://doi.org/10.1056/NEJMr1010517>.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, *337*, 816–821.
- Kumar, P., Gupta, V., Mishra, A., Modi, D., & Pandey, B. (2009). Potential of molecular markers in plant biotechnology. *Plant Omics*, *2*(4), 141–162.
- Kumar, S., Hamweih, A., Manickavelu, A., Kumar, J., Sharma, T. R., & Baum, M. (2014). Advances in lentil genomics. In S. Gupta, N. Nadarajan, & D. S. Gupta (Eds.), *Legumes in Omics Era* (pp. 111–130). New York: Springer Science + Business Media.
- Kumar, S., Rajendran, K., Kumar, J., Hamwieh, A., & Baul, M. (2015). Current knowledge in lentil genomics and its application for crop improvement. *Frontier in Plant Science*, *6*(78), 1–13.
- Lander, E. S. (1996). The new genomics; global views of biology. *Science (New York, N.Y.)*, *274*(5287), 536–539. Available from <https://doi.org/10.1126/science.274.5287.536>.
- Lu, Y. F., Goldstein, D. B., Angrist, M., & Cavalleri, G. (2014). Personalized medicine and human genetic diversity. *Cold Spring Harbor Perspectives in Medicine*, *4*(9), a008581. Available from <https://doi.org/10.1101/cshperspect.a008581>.
- Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., . . . Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, *339*, 823–826.
- Mardis, E. R. (2008). Next generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402. Available from <https://doi.org/10.1146/annurev.genom.9.081307.164359>.
- Marsden, R. L., Lewis, T. A., & Orengo, C. A. (2007). Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics*, *8*, 86. Available from <https://doi.org/10.1186/1471-2105-8-86>.
- McGrath, S., & van Sinderen, D. (Eds.), (2007). *Bacteriophage: Genetics and Molecular Biology* (1st ed.). Caister Academic Press, ISBN 978-1-904455-14-1.
- Mckusick, V. A. (2005). The Gordon Wilson lecture: The clinical Legacy of Jonathan Hutchinson (1828–1913): syndromology and dysmorphology meet genomics. *Transactions of the American Clinical and Climatological Association*, *116*, 15–38.
- Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Research*, *15*, 1767–1776. Available from <https://doi.org/10.1101/gr.3770505>.
- Min Jou, W., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, *237*(5350), 82–88. Available from <https://doi.org/10.1038/237082a0>.
- Mochida, K., & Shinozaki, K. (2010). Genomics and bioinformatics resources for crop improvement. *Plant & Cell Physiology*, *51*(4), 497–523. Available from <https://doi.org/10.1093/pcp/pcq027>.

- Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., & O'Neal, C. (1965). RNA code words and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, 53(5), 1161–1168. Available from <https://doi.org/10.1073/pnas.53.5.1161>.
- O'Donnell, C. J., & Nabel, E. G. (2011). Genomics of cardiovascular disease. *The New England Journal of Medicine*, 365(22), 2098–2109. Available from <https://doi.org/10.1056/NEJMra1105239>.
- Ogbe, R. J., Ochalefu, D. O., & Olaniru, O. B. (2016). Bioinformatics advances in genomics—a review. *International Journal of Current Research and Review*, 8(10), 5–10.
- Callinan, P. A., & Feinberg, A. P. (2006). The emerging science of epigenomics. *Human Molecular Genetics*, 15, 95–101. Available from <https://doi.org/10.1093/hmg/ddl095>.
- Pevsner, J. (2009). In N. J. Hoboken (Ed.), *Bioinformatics and functional genomics*. Wiley-Blackwell, ISBN 978-0-470-08585-1.
- Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, 10(4), 354–366. Available from <https://doi.org/10.1093/bib/bbp026>.
- Rahimi-Vahed, A. R., Rabbani, M., Tavakkoli-Moghaddam, R., Torabi, S. A., & Jolai, F. (2007). A multi-objective scatter search for a mixed-model assembly line sequencing problem. *Advanced Engineering Informatics*, 21(1), 85–99. Available from <https://doi.org/10.1016/j.aei.2006.09.007>.
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3), 441–448. Available from [https://doi.org/10.1016/0022-2836\(75\)90213-2](https://doi.org/10.1016/0022-2836(75)90213-2).
- Schuler, M. A., & Reichhart, D. W. (2003). Functional genomics of P450S. *Annual Review of Plant Biology*, 54, 629–667. Available from <https://doi.org/10.1146/annurev.arplant.54.031902.134840>.
- Scott, M. F., Ladejobi, O., Amer, S., et al. (2020). Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity*, 125, 396–416. Available from <https://doi.org/10.1038/s41437-020-0336-6>.
- Shendure, J., Balasubramanian, S., Church, G., et al. (2017). DNA sequencing at 40: past, present and future. *Nature*, 550, 345–353. Available from <https://doi.org/10.1038/nature24286>.
- Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26, 1135–1145. Available from <https://doi.org/10.1038/nbt1486>.
- Shi, Y., & Anderson, R. C. (2003). High-resolution single-stranded DNA analysis on 4.5 cm plastic electrophoretic microchannels. *Electrophoresis*, 24, 3371–3377.
- Shukla, V. K., Doyon, Y., Miller, J. C., DeKolver, R. C., Moehle, E. A., Worden, S. E., . . . Urnov, F. D. (2009). Precise genome modification in the crop species *Zea mays* using zinc-finger nucleases. *Nature*, 459, 437–441.
- Staden, R. (1979). A strategy of DNA sequencing employing computer program. *Nucleic Acids Research*, 6(7), 2601–2610. Available from <https://doi.org/10.1093/nar/6.7.2601>.
- Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews. Genetics*, 2(7), 493–503. Available from <https://doi.org/10.1038/35080529>.
- Tao, Y., Zhao, X., Mace, E., Henry, R., & Jordan, D. (2019). Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant*, 12, 156–169.
- ten Bosch, J. R., & Grody, W. W. (2008). Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *The Journal of Molecular Diagnostics*, 10(6), 484–492. Available from <https://doi.org/10.2353/jmoldx.2008.080027>.
- The ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science (New York, N.Y.)*, 306, 636–640.
- Thiel, T., Michalek, W., Varshney, R., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR markers in barley (*Hordeum vulgare* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 106, 411–422.
- Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively parallel sequencing: The next big thing in genetic medicine. *American Journal of Human Genetics*, 85(2), 142–154. Available from <https://doi.org/10.1016/j.ajhg.2009.06.022>.
- Uauy, C., Paraiso, F., Colasuonno, P., Tran, R. K., Tsai, H., Berardi, S., . . . Dubcovsky, J. (2009). A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. *BMC Plant Biology*, 9, 115.
- Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology*, 23, 48–55. Available from <https://doi.org/10.1016/j.tibtech.2004.11.005>.
- Venter, J. C., Adams, M. D., Sutton, G. G., et al. (1998). Shotgun sequencing of the human genome. *Science (New York, N.Y.)*, 280(5369), 1540–1542. Available from <https://doi.org/10.1126/science.280.5369.1540>.
- Watson, A., Ghosh, S., Williams, M. J., et al. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature Plants*, 4, 23–29. Available from <https://doi.org/10.1038/s41477-017-0083-8>.
- Xu, S. S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics*, 163, 789–801.
- Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., . . . Dubcovsky, J. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc Natl Acad Sci USA*, 103, 19581–19586.
- Zamir, D., & Ladizinsky, G. (1984). Genetics of allozyme variants and linkage groups in lentil. *Euphytica*, 33, 329–336. Available from <https://doi.org/10.1007/bf00021129>.
- Zhu, H., Choi, H. K., Cook, D. R., & Shoemaker, R. C. (2005). Bridging model and crop legumes through comparative genomics. *Plant Physiology*, 137, 1189–1196. Available from <https://doi.org/10.1104/pp.104.058891>.

This page intentionally left blank

Genome-wide predictions, structural and functional annotations of plant transcription factor gene families: a bioinformatics approach

Sudhanshu Srivastava¹, Kapil Gupta², Kanchan Yadav¹, Manoj Kumar Yadav³ and Dinesh Yadav¹

¹Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India, ²Department of Biotechnology, Siddharth University, Siddharthnagar, Uttar Pradesh, India, ³Department of Agricultural Biotechnology, College of Agriculture, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India

6.1 Transcription factor: an introduction

RNA polymerases, basal transcription proteins, transcription factors (TFs), coactivators, corepressors, and chromatin-related proteins such as histone acetyl transferases are key elements associated with gene expression and regulation at transcriptional level. TFs are sequence-specific DNA-binding proteins playing significant role in the regulation of diverse genes by targeting unique DNA sequences known as *cis*-elements present in the gene promoters. In general, TFs possess DNA-binding domain (DBD), an oligomerization motif, a transcription regulatory (activation) domain, and a nuclear localization signal. TFs show variability in DBDs by binding with specific DNA sequences present in the promoters and modulate temporal and spatial expression of specific genes. The activation domain, which is distinct from the DBD, is responsible for the combinatorial control of genes by a variety of TFs. TFs and regulatory elements interact and form the complexes with other bound TFs and facilitate RNA polymerase II recruitment to complex and start gene transcription. Application of next-generation sequencing techniques like transcriptome analysis and whole-genome sequencing has led to identifying several TF gene families and its members with the aid of tools of bioinformatics. The potential of these TFs gene families for crop improvement using biotechnological approach is emerging in the present era of omics. TFs account for about >7% of the coding sequences in plant genomes (Iida, Seki, Sakurai, & Satou, 2005; Udvardi, Kakar, Wandrey, & Montanari, 2007), and substantial efforts have been made to elucidate the functions of TFs in biological processes during the last two decades.

6.2 Plant transcription factors and its multifarious applications

Plant-specific TFs are involved with diverse functions associated with the growth and development of plants. It is associated with several functions like morphology, inflorescence/flower formation, reproduction, embryogenesis, fruit development and ripening, plant morphology, organ development, senescence, signaling, metabolism, and abiotic and biotic stress responses (Yadav, Malviya, Nasim, & Kumar, 2016). TF families such as APETALA2 (AP2)/ethylene-sensitive factor (ERF), basic-domain leucine zipper (bZIP), basic helix–loop–helix (bHLH), DNA binding with one finger (Dof), myeloblastosis (MYB), MADS, NAM/ATAF/CUC (NAC), WRKY, and zinc fingers are briefly discussed. These TF families have a significant number of members and play several essential roles exclusively observed in plants. Large numbers of TF family genes in plant genomes occurred due to a higher rate of expansion in specific TF families compared to those in other biological kingdoms (Yadav et al., 2016). This expansion in TF family genes has allowed functional divergence and acquisition of novel and adaptive roles

TABLE 6.1 Some of the important transcription factors with DNA-binding domains, probable DNA-binding sequence, and predicted three-dimensional (3D) models.

Sl. no.	Type of TF	No. of A.A. residues in DNA-binding domain	Probable DNA-binding sequence	Quick access to predicted 3D model
1.	bZIP	60	(T/G/C) ACGTG	https://www.rcsb.org/3d-view/6IAK
2.	bHLH	18	G-box DNA sequence motif (CACGTG)	https://www.rcsb.org/3d-view/5GNJ
3.	Zinc finger	50–80 (in C ₂ H ₂ type)	A(G/C)T repeat, TGCTANNATTG, TACAAT, A[AG/CT]-CNAC, etc. are possible binding sites	https://www.rcsb.org/3d-view/6JNN
4.	Trihelix (HLHLH)	70	Photo responsive GT element: 5'-G-Pu-(T/A)-A-A-(T/A)-3'	https://www.rcsb.org/3d-view/2JMW
5.	HMG box	75–80	AT-rich region in minor groove of DNA	https://www.rcsb.org/3d-view/1J5N
6.	Homeodomain (HD)	60	CAAT(A/T)AYYG or CAAT(G/C)AYYG	https://www.rcsb.org/3d-view/6ES3
7.	MADS-box	50–60	CA-rich G-box CC(AT) ₆ GG	https://www.rcsb.org/3d-view/4OX0
8.	MYB protein	Three copies of 53 AA repeat	AACNG, H-box: (CCTACC)	https://www.rcsb.org/3d-view/6KKS
9.	HSFs	7–21	HSE: nTTCnnGAA-nnTTCn	https://www.rcsb.org/3d-view/1FBS
10.	AP2/EREBP	68-AA repeat unit (in AP2) 59-AA (in EREBP)	GCC-box of EREBP gene	https://www.rcsb.org/3d-view/5WX9

HSE, Heat shock element; *EREBP*, ethylene responsive element binding protein.

in plants. In general, TFs are classified based on the unique DBD, and some of the important plant-specific TFs revealing the diversity of DBDs, DNA-binding sequences, and predicted three-dimensional (3D) structures are shown in Table 6.1. A brief introduction to some of the important plant TF gene families with major emphasis on bioinformatics studies is highlighted.

6.2.1 AP2/ERF family

The AP2/ERF TF family is one of the largest plant-specific TF groups comprising two subfamilies AP2 and ERF (ethylene-responsive element-binding factors). It has a highly conserved 58-amino-acid long AP2/ethylene-responsive element-binding domain. It is associated with several functions like flower development, cell proliferation, secondary metabolism, and abiotic and biotic stress responses (Agarwal, Gupta, Lopato, & Agarwal, 2017; Feng, Hou, Xing, & Liu, 2020; Xie, Nolan, Jiang, & Yin, 2019). Substantial bioinformatics-based structural and functional characterization of identified AP2/ERF gene families from sequenced genomes of many crops has been reported in recent years. A bioinformatics-based genome-wide mining of durum wheat genome revealed a total of 271 members of AP2/ERF genes, many of them associated with abiotic stresses (Faraji, Filiz, Kazemitabar, & Vannozzi, 2020). Similarly, in silico analysis of genomes of rice, *Brassica oleracea*, sunflower, Tartary buckwheat, sugarcane, and pineapple identified 170, 226, 288, 218, and 97 genes, respectively (Li, Chai, Lin, & Huang, 2020; Li, Fan, Yang, & Hu, 2020; Liu, Ma, Sun, & Huang, 2019; Liu, Sun, Ma, & Zheng, 2019; Najafi, Sorkheh, & Nasernakhaei, 2018; Rashid, Guangyuan, Guangxiao, Hussain, & Xu, 2012; Thamilarasan, Park, Jung, & Nou, 2014; Zhang, Pan, Liu, & Lin, 2021).

6.2.2 bHLH family

This family is prevalent among eukaryotes, including plants and animals, following an independent evolutionary event. The conserved domain is made up of 60 amino acids with an N-terminal stretch of 18 hydrophilic and essential amino acids associated with DNA binding and a helix–loop–helix (HLH) domain of hydrophobic residues separated by an intervening loop that forms homo- or heterodimers with interacting proteins (Feller, Machemer, Braun, & Grotewold, 2011; Fernández-Calvo, Chini, & Fernández-Barbero, 2011). These TFs are involved in diverse biological processes like metabolite biosynthesis, carpel and fruit development, suppression of seed germination cotyledon expansion, anthocyanin biosynthesis, growth, JA (Jasmonate)-mediated defense processes, and abiotic and biotic stresses. Genome-wide identification and characterization using bioinformatics tools have been attempted in several crops. An analysis of rice genome revealed 167 bHLH genes, and its comparison with *Arabidopsis* formed 25 subfamilies supported by phylogenetic tree (Li, Duan, Jiang, & Sun, 2006). Mining of genomes of wheat, ginseng, potato, and *Dendrobium officinale*, a traditional Chinese herb, identified 225, 169, 124, and 98 bHLH genes, respectively (Chu, Xiao, Su, & Liao, 2018; Guo & Wang, 2017; Wang & Liu, 2020; Wang, Zhao, Kong, & Lu, 2018). Using comparative genomic approach, functional analysis of 183 rice, 231 maize, and 571 wheat bHLH genes has also been reported using bioinformatics tools recently (Wei & Chen, 2018).

6.2.3 bZIP

bZIP TFs represent an important TF family ubiquitously found in all eukaryotes, and plants possess several family members with a high level of diversity (Riechmann, Heard, Martin, & Reuber, 2000). It comprises a conserved domain of 60–80 amino acids with two distinct regions, namely, a basic region followed by leucine zipper. The basic region contains the N-x7-R/K-x9 motif involved with DNA binding and nuclear localization, while leucine zipper is associated with homo- and heterodimerization (Jakoby, Weisshaar, Dröge-Laser, & Vicente-Carbajosa, 2002). bZIP TFs in *Arabidopsis* are categorized into ten classes based on structural and functional characteristics (Jakoby et al., 2002). Plant bZIPs bind to the (T/G/C) ACGTA *cis*-element preferentially; group A members, such as ABRE-binding factor, identify the ABRE (PyACGTGG/TC) *cis*-element and mediate ABA (abscisic acid)-mediated expression of stress-responsive genes. Diverse functions like organ growth, floral development, cell cycle, seed maturation and germination, photomorphogenesis, light signaling, and responses to stresses are regulated by this TF (Ali, Sarwat, Karim, & Faridi, 2016; Alves, Dadalto, Goncalves, & De Souza, 2013). Using bioinformatics approach, several sequenced plant genomes have been characterized for the presence of multiple bZIP genes. Genome mining of strawberry revealed 54 bZIP genes (Lu, Wang, Zhang, & Feng, 2020). Similarly, 50 and 45 bZIP genes identified from genomes of *Arachis duranensis* and *Arachis ipaensis* were functionally characterized for elucidating its role in seed development and response to salt stress (Wang, Yan, Wan, & Huai, 2019; Wang, Zhang, Hu, & Guo, 2019). A total of 191 bZIP genes identified from genome of wheat were analyzed for its role in abiotic stress tolerance recently (Agarwal, Baranwal, & Khurana, 2019). Similarly, genome-wide bioinformatics analysis of celery, an important vegetable revealed 62 bZIP genes that were functionally characterized for its role in abiotic stresses (Yang, Feng, Xu, & Duan, 2019).

6.2.4 DNA binding with one finger family

It is an important plant-specific TF with 50–52 amino acids long conserved DBD at the N-terminus and a transcriptional regulatory domain at the C-terminus. The CX₂CX₂1CX₂C motif of DBD has been predicted to form a single zinc finger mediated by four conserved cysteine (Cys) residues and belongs to C₂C₂ Zn finger family, and, therefore, named Dof domain proteins (Kushwaha, Gupta, Singh, Rastogi, & Yadav, 2011; Yanagisawa, 2002). Four Cys residues in the DBD mediate sequence-specific identification of the *cis*-element and the variable C-terminal, which is responsible for di- or oligomerization and protein–protein interactions (Yanagisawa et al., 2002; Gupta, Malviya, Kushwaha, & Nasim, 2015; Gupta, Arya, Malviya, Bisht, & Yadav, 2016). Dof TFs recognize the AAAG or CTTT sequences located in the promoters of Dof target genes. It is associated with several functions like seed storage protein accumulation, seed dormancy, photosynthetic control, flowering, phytohormone reaction, and biotic and abiotic stress responses (Gupta et al., 2016; Noguerro, Atif, Ochatt, & Thompson, 2013).

Attempts have been made for in silico prediction of *Dof* gene families for genome-sequenced crops. The variable number of *Dof* genes reported in various plant species, namely, *Hordeum vulgare*, *Triticum aestivum*, *Sorghum bicolor*, *Zea mays*, *Brachypodium distachyon*, *Solanum lycopersicum* and *Saccharum officinarum*, *Chinese cabbage*, *Vitis vinifera*, *Phaseolus vulgaris*, and *Solanum melongena* is 24, 31, 28, 54, 27, 34, 25, 76, 25, 36, and 29, respectively

(Cai, Zhang, Zhang, & Zhang, 2013; da Silva, da Silveira Falavigna, Fasoli, & Buffon, 2016; Gupta, Kushwaha, Singh, & Bisht, 2014; Hernando-Amado, González-Calle, Carbonero, & Barrero-Sicilia, 2012; Jiang, Zeng, Zhao, & Zhang, 2012; Kushwaha et al., 2011; Ma, Li, Wang, Tang, & Xiong, 2015; Moreno-Risueno, Martinez, Vicente-Carbajosa, & Carbonero, 2007; Shaw, McIntyre, Gresshoff, & Xue, 2009; Wei et al., 2018).

6.2.5 MADS family

The MADS TF family is widely observed in eukaryotes and possesses a conserved MADS DBD. The name MADS was coined using the initials from first four identified members of this family; MAINTENANCE OF MINICHROMOSOME1 (*Saccharomyces cerevisiae*), AGAMOUS (*Arabidopsis thaliana*), DEFICIENS (*Antirrhinum majus*), and Serum Response Factor (*Homo sapiens*). The family of MADS-box TFs expanded in flowering plants during their evolution. The MADS box genes possess a highly conserved 55–60 amino acids long DBD, named MADS domain at N-terminal. MADS TFs recognize the 10-base-pair AT-rich motif CARG-box as dimers (Kappel, Eggeling, Rumppler, & Groth, 2021). This TF plays an important role in developmental regulations, changes in vegetative phase, specification of floral organs, ovule and female gametophyte development, fruit ripening, and root growth (Muino, Smaczniak, Angenent, Kaufmann, & van Dijk, 2014). Genome-wide studies using bioinformatics approach have been reported in several plants in recent years. In rice a total of 75 MADS box genes have been identified, and its involvement in reproductive development and stress has been attempted (Arora, Agarwal, Ray, & Singh, 2007). A total of 131 and 91 MADS-box genes were predicted from genomes of *S. lycopersicum* and *B. oleracea* genomes and were characterized for elucidating its role in floral development (Sheng, Zhao, Wang, & Yu, 2019; Wang, Yan, et al., 2019; Wang, Zhang, et al., 2019). Similarly, genome mining of lotus genome revealed 44 genes, many of which were involved in floral development (Lin, Cao, Damaris, & Yang, 2020). Genome analysis of pomegranate identified 36 MIKC-type MADS box genes, and its role in the peel and inner seed coat development was reported recently (Zhao, Wu, Zhang, & Wang, 2020; Zhao, Ye, Wang, Wang, & Chen, 2020; Zhao, Zhao, Wang, & Zhang, 2020).

6.2.6 Myeloblastosis family

The MYB TF is ubiquitously found in most of the eukaryotes, including plants. The DBD, known as MYB repeat, consists of two helices bound by a turn known as the helix-turn-helix motif. The domain has one to four conserved MYB repeats (R) of 52 amino acids at N-terminus (Jia, Tong, & Wang, 2004). This TF family on the basis of number and position of repeats has four groups, 1R-MYB, R2R3-MYB, R1R2R3-MYB, and 4RMYB (Dubos, Stracke, Grotewold, & Weisshaar, 2010). This family serves many functions in plants, including primary and secondary metabolism, plant growth, leaf polarity, trichome development, cell fate, cell wall biogenesis, hormone signal transduction, and abiotic and biotic stress responses (Baldoni, Genga, & Cominelli, 2015; Dubos et al., 2010; Li, Xiong, Li, & Ye, 2019). In silico genome-wide identification and characterization of MYB family genes from several sequenced plant genomes have been reported. A comparative genome-wide identified 155 and 197 MYB genes from rice and Arabidopsis has been substantially characterized using bioinformatics tools for several attributes (Katiyar, Smita, Lenka, & Rajwanshi, 2012). Genome mining of cotton revealed 524 genes, many of which were associated with fiber development (Salih, Gong, He, & Sun, 2016). Similarly, genome-wide studies of potato and *Physcomitrella patens* identified 158 and 116 MYB genes, respectively (Pu, Yang, Liu, & Dong, 2020; Sun, Ma, Chen, & Liu, 2019). A genome-wide comparative analysis of *Musa acuminata* and *Musa balbisiana* genomes revealed 305 and 252 MYB genes, respectively. Its role in fruit ripening was also elucidated (Tan, Ijaz, Salih, & Cheng, 2020).

6.2.7 NAM/ATAF/CUC family

NAC is an important plant-specific TF having NAC domain of 150 amino acids with 5 subdomains designated as (A–E) at N-terminus and a variable C-terminal domain. The NAC DBDs undergo dimerization and form a central semi- β -barrel with seven twisted antiparallel β -strands along with three α -helices and strands of β -sheets are associated with DNA-binding function (Chen, Wang, Xiong, & Lou, 2011). It has been observed that few NAC TFs have more than one NAC domain and nuclear localization signals as monopartite, bipartite, or multipartite. It encodes up to 45 chimeric proteins, including WRKY, TIR, LRR, protein kinase, peptidase A1, DNAJ, ZF B, and other domains, which might enable them to control a complex interacting network (Mohanta, Yadav, Khan, &

Hashem, 2020). NAC TFs could function as both a TF and an enzyme. The ABI, MYB, DREB2, WRKY, JUMONJI, and KNAT TFs interact with the NAC TFs (Mohanta et al., 2020). It is involved with several plant-specific functions like flowering, anther dehiscence, lateral root growth, and biotic and abiotic stresses (Agarwal, Shukla, Gupta, & Jha, 2013). Genome-wide bioinformatics assessment of genomes of *Medicago truncatula* and *Medicago sativa* revealed 97 and 113 NAC genes, respectively (Ling, Song, Wang, & Guo, 2017; Min, Jin, Zhang, & Wei, 2020). Similarly, genome mining of pepper identified 104 genes (Diao, Snyder, Wang, & Liu, 2018). In wheat a total of 488 NAC genes were identified from the genome and in silico characterized for its role in drought and heat stresses (Guerin, Roche, Allard, & Ravel, 2019). A total of 102 NAC genes were identified from genome of cocoa tree, a major beverage crop recently (Shen, Zhang, Shi, & Sun, 2019). In a genome-wide analysis of Tartary buckwheat, a medicinal plant showed 80 NAC genes, and many of them were associated with fruit development (Liu, Ma, et al., 2019; Liu, Sun, et al., 2019).

6.2.8 WRKY family

It has a conserved DBD at N-terminal with WRKYGQK motif and a putative zinc finger motif at its carboxyl terminal and recognizes the W-box (C/T) TGAC(T/C) of target genes predominately associated with abiotic stress (Chen et al., 2012; Duan, Nan, Liang, & Mao, 2007). Plant growth and development, trichome development, embryogenesis, seed coat, secondary metabolism, and abiotic and biotic stress tolerance are all controlled by WRKY TFs (Wei, Chen, Chen, Wu, & Xie, 2012). Genome-wide identification of WRKY genes from genomes of several crops revealed diversity in terms of variability in the number of genes. Genome mining of lotus and cultivated strawberry identified 65 and 47 WRKY genes (Chen & Liu, 2019; Li et al., 2019). Bioinformatics-based identification and functional characterization under abiotic stresses in potato, chickpea, and buckwheat of 79, 70, and 78 WRKY genes, respectively, have been reported recently (He, Li, Chen, & Yang, 2019; Waqas, Azhar, Rana, & Azeem, 2019; Zhang, Wang, Yang, & Kong, 2017). Genome mining of wheat genome for WRKY genes and its characterization exclusively for its role in abiotic stress have been attempted (Gupta, Mishra, Kumari, & Raavi, 2019). In cucumber and *Artemisia annua*, a total of 61 and 122 WRKY genes were identified from the genome using bioinformatics approach (Chen, Chen, Han, & Lu, 2020; De Paolis, Caretto, Quarta, & Di Sansebastiano, 2020).

6.2.9 Zinc fingers

Zinc finger protein (ZFP) TF family is a predominant eukaryotic TFs family first reported from *Xenopus* oocytes as TFIIIA (Miller, McLachlan, & Klug, 1985). Plants, unlike animals, seem to have evolved to various specialized roles by adapting conventional zinc-finger motifs and evolving new zinc finger domains. ZFPs possess a zinc finger motif that binds to DNA by means of Cys and histidine (His) residues and forms finger-like projections that bind the DNA. C2H2-type ZFPs are the well-studied ZFPs in plants, with a DBD of 30 amino acids with 2 conserved Cys and His residues attached to 1 zinc ion tetrahedrally (CX2–4CX3FX5-LX2HX3–5H). Some zinc finger motifs, such as GATA and PHD, are associated with DBD of TFs, while others, such as LIM and RING-finger, mediate protein–protein interactions. Other motifs with zinc fingers, such as Dof and WRKY, have been held as separate families due to the unusual spacing between cytosines in Dof and Zn chelating residues in WRKY, as well as the absence of two hydrophobic amino acids (F and L) present in the C2H2/TFIIIA type. It is associated with several functions influencing plant growth and development, plant architecture, trichome development, hormone signaling, and stress responses (Liu, Liu, Hu, & Hua, 2017).

In the recent years with the availability of genome sequences, several bioinformatics studies targeting on genome-wide identification and characterization of ZFPs TF from several crops have been reported. Using comparative genomics approach, attempts have been made to characterize 68 and 67 CCCH-type zinc finger TF genes of Arabidopsis and rice, respectively (Wang, Guo, Wu, & Yang, 2008). Genome mining of chickpea identified 58 CCCH-type zinc finger TF genes, many of which were associated with abiotic stress (Pradhan, Kant, Verma, & Bhatia, 2017). In grapevine a total of 98 C2C2 zinc finger TF genes were analyzed by genome mining, and its role in pollen development was studied (Arrey-Sales, Caris-Maldonado, Hernandez-Rojas, & Gonzales, 2021). Genome-wide study of *M. truncatula* revealed 218 C2H2-type zinc finger TF genes (Jiao, Wang, Du, & Wang, 2020), while 118 genes were identified from genome of tobacco (Yang, Chao, Wang, & Hu, 2016). Using RNA-seq data, a total of 32 C2H2-type zinc finger TF genes were characterized from tomato, and its role in biotic and abiotic stress was deciphered (Zhao, Wu, et al., 2020; Zhao, Ye, et al., 2020; Zhao, Zhao, et al., 2020).

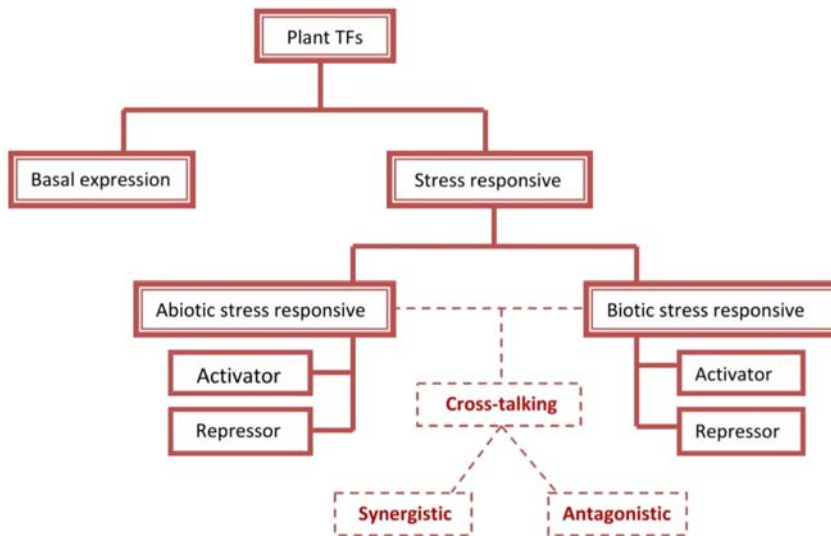


FIGURE 6.1 Schematic representation of plant-specific TF showing stress responsiveness. *TF*, Transcription factor.

A comparative analysis of 386, 196, and 195 C2H2-type zinc finger TF genes of *Gossypium hirsutum*, *Gossypium arboreum*, and *Gossypium raimondii*, respectively, has been attempted using bioinformatics approach to reveal its role in fiber development (Salih, Odongo, Gong, He, & Du, 2019).

6.3 Transcription factors for biotic and abiotic tolerance

The biotechnological approaches for crop improvement mainly target toward the development of stress-tolerant crops as huge losses due to several biotic and abiotic stresses are observed in several crops. Plant TFs have great potential to develop stress-tolerant crops, and these special classes of TFs are also referred to as stress responsive, which may show synergistic or antagonistic effect based on the regulation of activator or repressor elements as shown in Fig. 6.1. The involvement of TF for developing stress-tolerant crops is supported by some examples as shown in Table 6.2.

6.4 Transcription factor databases

In the recent years, several sequenced plant genomes were annotated for the presence of different types of TFs using bioinformatics tools based on the presence of unique DBD sequences (Riechmann et al., 2000). Crop-specific databases have been created for the benefit of the researchers like Arabidopsis genome sequencing initially revealed more than 1500 TFs representing 30 different TF families (Riechmann et al., 2000). Some of the important plant-specific TF databases are listed in Table 6.3.

6.5 Bioinformatics tools used for structural and functional analysis of transcription factor gene families

For the characterization of TF gene families, bioinformatics tools targeting for genome-wide identification, phylogenetic, 3D structural prediction, and validation and in silico expression profiling have been developed (Fig. 6.2). The steps of major bioinformatics analysis are described in Flow Charts 1, 2, 3 and 4.

TABLE 6.2 Important plant transcription factors associated with biotic and abiotic stress tolerance.

Sl. no.	TF gene family	Stress type	Examples	Stress	Target plant	References
1.	bZIP	Biotic	StbZIP61	<i>Phytophthora infestans</i>	<i>Solanum tuberosum</i>	Zhao, Wu et al. (2020), Zhao, Ye et al. (2020), Zhao, Zhao et al. (2020)
		Abiotic	GmbZIP44	Salinity, freezing	<i>Arabidopsis thaliana</i>	Liao, Zou, Wei, and Hao (2008)
2.	bHLH	Biotic	OsHLH96	Brown plant hopper	<i>Oryza sativa</i>	Wang, Yan et al. (2019), Wang, Zhang et al. (2019)
		Abiotic	AtMYC2	Osmotic stress	<i>A. thaliana</i>	Abe et al. (2003)
3.	Zinc finger	Biotic	VvZFP11	S.A., methyl jasmonate, <i>Erysiphe necator</i>	<i>Vitis vinifera</i>	Yu, Li, and Wu (2016)
		Abiotic	Alfin1	Salinity	<i>Alfalfa</i>	Bastola, Pethe, and Winicov (1998)
			CaZF	Salinity	<i>Tobacco</i>	Jain, Roy, and Chattopadhyay (2009)
			OSISAP2	Freezing		Xu and Cui (2007)
		ZPT2–3	Drought	<i>Petunia</i>	Sugano, Kaminaka, Rybka, and Catala (2003)	
		ZAT7	Salinity	<i>Arabidopsis</i>	Ciftci-Yilmaz, Morsy, Song, and Coutu (2007)	
ZAT12	Light stress	Davletova, Schlauch, Coutu, and Mittler (2005)				
4.	Trihelix (HLHLH)	Biotic	OsRML1	<i>Magnaporthe grisea</i>	<i>O. sativa</i>	Wang, Chen et al. (2004), Wang, Hong et al. (2004)
		Abiotic	GmGT-2A and 2B	Salt, drought, and freezing	<i>Glycine max</i>	Xie, Zou, Lei, and Wei (2009)
5.	WRKY	Biotic	CsWRKY50	<i>Pseudoperonospora cubensis</i>	<i>Cucumis sativus</i>	Luan, Chen, Liu, and Li (2019)
		Abiotic	GmWRKY21	Freezing	<i>A. thaliana</i>	Zhou, Tian, Zou, and Xie (2008)
6.	Homeodomain (HD)	Biotic	StWRKY1	<i>P. infestans</i>	<i>S. tuberosum</i>	Yogendra, Kumar, Sarkar, and Li (2015)
		Abiotic	GmPHD2	Salt stress	<i>Arabidopsis (transgenic)</i>	Wei, Huang, Hao, and Zou (2009)
7.	NAC	Biotic	OsNAC6	<i>M. grisea</i>	<i>O. sativa</i>	Nakashima, Tran, Nguyen, and Fujita (2007)
		Abiotic	AtNAC2 AtNAC019 AtNAC055	Drought	<i>A. thaliana</i>	Tran, Nakashima, Sakuma, and Simpson (2004)
			ONAC045 SNAC1	Drought, salinity	<i>O. sativa</i>	Zheng, Chen, Lu, and Han (2009)
8.	MYB protein	Biotic	AtMYB102	GPA	<i>A. thaliana</i>	Zhu, Guo, Ma, Wang, and Zhou (2018)
			TaRIM1	<i>Rhizoctonia cerealis</i>	<i>Triticum aestivum</i>	Shan, Rong, Xu, and Du (2016)
		Abiotic	OsMYB4	Drought	<i>Lycopersicon esculentum</i>	Vannini, Locatelli, Bracale, and Magnani (2004)
9.	AP2/ERF	Biotic	TaPIEP1	<i>Bipolaris sorokiniana</i>	<i>T. aestivum</i>	Dong, Liu, Lu, and Du (2010)
		Abiotic	OsDRAP1	Drought	<i>O. sativa</i>	Huang, Wang, Wang, and Zhao (2018)
			ZmERFB180	Waterlogging	<i>Zea maize</i>	Yu, Liang, Fang, and Zhao (2019)

GPA, Green peach aphid.

TABLE 6.3 List of plant-specific transcription databases.

Sl. no.	Database name	Basic features	References
1.	Phytozome	<ul style="list-style-type: none"> • It is a platform of comparative genomics. • v11.0 gives access to 65 sequenced and annotated genomes of green plants. • v12.0 contains 93 assembled and annotated genomes from 82 <i>Viridiplantae</i> species. • v13 is the latest, known as “Phytozome-Next” that contains 224 assembled and annotated genomes from 128 <i>Archaeplastida</i> spp. • This new version comes with the capability of “clade-cutting.” • Web: https://phytozome-next.jgi.doe.gov/ 	David et al. (2012)
2.	PlantTFDB	<ul style="list-style-type: none"> • This TF database of plants contains information related to different species with and without their genome sequences, TF family, number, annotation, orthologous group, phylogenetic trees, etc. • v4.0 presents a collection of TFs. for 165 plant species. TF prediction server is upgraded for previous TF families, and four new tools are introduced for regulation prediction and functional enrichment analysis. • An updated annotation for the TFs as well as TFex module (extended TF repertoire) of newly sequenced species have been added in v5.0. • Web: http://planttfdb.gao-lab.org/ 	Jin, Tian, Yang, and Meng (2017)
3.	PlnTFDB	<ul style="list-style-type: none"> • It is plant TF database that provides a basic description for each TF family from literature reference. • New version (3.0) provides a complete set of putative TFs, genome of which is totally sequenced. • There are more rules set up for the classification of transcription factors. • 16 more plant families have been added in v3.0. • Web: http://plntfdb.bio.uni-potsdam.de/v3.0/ 	Perez-Rodriguez, Riano-Pachon, Correa, and Rensing (2009)
4.	JASPAR	<ul style="list-style-type: none"> • It provides full accessibility to carefully arranged eukaryotic TFs, their binding profile and their TF flexible models across six taxonomic groups, vertebrata, nematode, insect, plantae, fungi, and urochordata. • JASPAR-2016: released with expanded CORE collection of 494 new TFBP of which 164 for plants. • JASPAR-2018 (seventh release): comes with 322 new PFMs of which 262 for plants. • JASPAR-2020 (eighth release): the latest version, introduced with updated CORE collection, includes 245 new PFMs/TFBPs of which 42 for plants. • Web: http://jaspar.genereg.net/ 	Fornes, Castro-Mondragon, and Khan (2019)
5.	TRRD	<ul style="list-style-type: none"> • It is transcription regulatory region database and generally incorporates only experimentally confirmed structure. • The regulatory units included in this database are <i>cis</i>-acting DNA elements, composite elements, promoters, enhancer, silencers, etc. • Website “http://www.mgs.bionet.nsc.ru/mgs/gnw/trrd/whats_new.shtml” contains no new information about any latest release after v7.0, September 2005. • v7.0 released with a new feature “TRRDSTARTS” that contains information of experimentally determined transcription start site. 	Kolchanov, Ignatieva, Ananko, and Podkolodnaya (2002)
6.	PlantCARE	<ul style="list-style-type: none"> • It is a database associated with plant <i>cis</i>-acting regulatory elements. It is generally used for the analysis of promoter sequences. • Web: http://bioinformatics.psb.ugent.be/webtools/plantcare/html/ 	Lescot, Hais, Thijs, and Marchal (2002)
7.	HEATSTER	<ul style="list-style-type: none"> • Heat stress—responsive TFs database. • Web: https://applbio.biologie.uni-frankfurt.de/hsf/heatster/ 	Scharf, Berberich, Ebersberger, and Nover (2012)

(Continued)

TABLE 6.3 (Continued)

Sl. no.	Database name	Basic features	References
8.	GreenPhyl	<ul style="list-style-type: none"> This database is created to assist the comparative functional genomics in plants species (from algae to higher). Version 5.0 possesses a list of gene families with 19 pangenomes (e.g., rice, maize, banana, grape, and cacao) and 27 reference genomes of around 46 species. Web: https://www.greenphyl.org/cgi-bin/index.cgi 	Guignon, Toure, Droc, and Dufayard (2020)
9.	LegumeTFDB	<ul style="list-style-type: none"> This database makes users accessible to specific collection of TFs of three legumes, <i>Glycine max</i>, <i>Lotus Japonicus</i>, and <i>Medicago truncatula</i>. The database holds information related to TF genes, sequence features, promoters, Gene Ontology assignment, etc. Web: http://legumetfdb.psc.riken.jp/ 	Mochida, Yoshida, Sakurai, and Yamaguchi-Shinozaki (2010)
10.	PvTFDB	<ul style="list-style-type: none"> This is the database of TFs of common bean, that is, <i>Phaseolus vulgaris</i>. It came into existence after the publication of common bean genome sequence. Reviews on this database assert that it holds information regarding 49 TF families classified from predicted set of 2370 TFs. Web: http://www.multiomics.in/PvTFDB/ 	Bhawna, Bonthala, and Gajula (2016)
11.	SoybeanTFDB	<ul style="list-style-type: none"> It is constructed by computationally analyzing the genome sequence data of soybean. A total of 61 TF families have been classified by identifying 4342 gene loci responsible to encode 5035 TF models. Now this Japanese database has been closed. Web: http://soybeantfdb.psc.riken.jp/index.pl 	Mochida, Yoshida, Sakurai, and Yamaguchi-Shinozaki (2009)
12.	SoyDB	<ul style="list-style-type: none"> It is a database of soybean TFs, which was prepared by the Dept. of Energy-Joint Genome Institute (DOE-JGI). It holds information related to protein sequences, predicted tertiary structures, putative DNA-binding sites, domains, etc. 	Wang, Libault, and Joshi (2010)
13.	CicerTransDB	<ul style="list-style-type: none"> It is the TF database of Chickpea made by NIPGR, India to facilitate the comprehensive study of TFs in this genus. The developers classified 1124 TFs of chickpea into 47 families. This platform is not limited to sequences of genes, proteins, and promoters rather it provides accessibility to motifs, domains, Gene Ontology, and homologs in PlantTFDB as well as TAIR. Studies carried out to develop this database explored 68 more TFs in chickpea, which was not reported previously in PlantTFDB. Web: http://www.cicertransdb.esy.es/index.html 	Gayali, Acharya, and Lande (2016)
14.	PpTFDB	<ul style="list-style-type: none"> Pigeonpea transcription factor database. Web: http://14.139.229.199/PpTFDB/Home.aspx 	Singh, Sharma, Singh, and Sharma (2017)
15.	GRASSIUS	<ul style="list-style-type: none"> It is publicly available online resource created by the integration of databases (GrassTFDB) plus computational and experimental datasets. GrassTFDB is an all-inclusive compilation of MaizeTFDB; RiceTFDB; SorghumTFDB; SugarcaneTFDB; and BrachypodiumTFDB. They plan to include more grasses in their databases, when sequence information of these grasses will become available. Apart from TF database, they are trying to construct TF ORFome collection (under development). Web: https://grassius.org/ 	Yang, Li, Jiang, and Yu (2017)
16.	RiceSRTFDB	<ul style="list-style-type: none"> This database gives expression information of TFs involved in stress conditions and various developmental processes in rice. Web: http://www.nipgr.res/RiceSRTFDB.html 	Priya and Jain (2013)

(Continued)

TABLE 6.3 (Continued)

Sl. no.	Database name	Basic features	References
17.	RicetissueTFDB	<ul style="list-style-type: none"> This database provides expression information of TFs of rice but in a tissue-specific manner. It contains 59 families of TFs classified by validating 1087 TFs after the analysis of 3078 TFs. Web: http://221.237.158.212.50009/index 	Chen, Chen, Luo, and Liao (2019)
18.	GramineaeTFDB	<ul style="list-style-type: none"> It is portal for comparative and functional genomics which includes all putative TFs from six grass species. Each TF contains specific details of sequence feature, promoter region, domain alignment, GO assignment, etc. Users can search putative <i>cis</i>-elements of promoter sites of TFs whose genome is sequenced. The given hyperlinks make users capable of accessing the expression profile of TF genes in maize, rice, and barley. Web: http://gramineaeetfdb.psc.riken.jp 	Mochida, Yoshida, Sakurai, and Yamaguchi-Shinozaki (2011)
19.	RED (Rice expression database)	<ul style="list-style-type: none"> This Chinese database is a collection of gene expression profile obtained from RNA sequence data of 284 high-quality RNA-Seq experiments. The v2.0 comes with a new annotation system IC4R-Seq. Web: http://expression.ic4r.org/ 	Xia, Zou, Sang, and Xu (2017)
20.	RiceXPro	<ul style="list-style-type: none"> Rice Expression profile (RiceXPro) database is a collection of gene expression profiles obtained from microarray of rice plant tissues grown in natural field conditions. Web: https://ricexpro.dna.affrc.go.jp/ 	Sato, Takehisa, Kamatsuki, and Minami (2013)
21.	DRTF	<ul style="list-style-type: none"> Database of rice TFs contains putative TFs of <i>O. sativa</i> (subspecies <i>indica</i> and <i>japonica</i>) scattered in 63 families. Web: http://drtf.pku.edu.cn 	Gao, Zhong, Guo, and Zhu (2006)
22.	CamRegBase	<ul style="list-style-type: none"> Camelina gene regulation database is the collection of RNA-Seq experiment data. v1. Includes collection of TFs and coactivators of Camelina. Web: https://camregbase.org/ 	Gomez-Cano, Carey, Lucas, and García Navarrete (2020)
23.	AGRIS	<ul style="list-style-type: none"> Arabidopsis Gene Regulatory Information Server (AGRIS)-2019 comprises three databases: <i>AtcisDB</i>: includes about 33,000 upstream regions of annotated genes. <i>AtTFDB</i>: database of 1770 TFs. <i>AtRegNet</i>: information related to 1,638,778 interactions between promoter and TF. Web: https://agris-knowledgebase.org/ 	Yilmaz, Mejia-Guerra, Kurz, and Liang (2011)
24.	ATR (Sheen lab)	<ul style="list-style-type: none"> Arabidopsis transcription regulators database is a classified collection of TF gene family. Web: http://genetics.mgh.harvard.edu/sheenweb/AraTRs.html 	Yoo, Cho, and Sheen (2007)
25.	wDBTF	<ul style="list-style-type: none"> This wheat TF database contains 40 families and 84 subfamilies created by analyzing 7112 gene sequences (contigs and singletons) of wheat. Web: http://www.appli.nantes.inra.fr:8180/wDBFT/ 	Romeuf, Tessier, and Dardevet (2010)
26.	TreeTFDB	<ul style="list-style-type: none"> This database contains the TF repertoires of six plant species, <i>Jatropha curcas</i>, <i>Carica papaya</i>, <i>Manihot esculenta</i>, <i>Populus spp.</i>, <i>Ricinus communis</i>, and <i>Vitis vinifera</i>. Main features: sequences, domain alignment, Gene Ontology assignment, etc. Additional features: full-length cDNAs, <i>cis</i>-motifs located in promoter, their id, and positions. Web: http://treetfdb.bmep.riken.jp/index.pl 	Mochida, Yoshida, Sakurai, and Yamaguchi-Shinozaki (2013)

(Continued)

TABLE 6.3 (Continued)

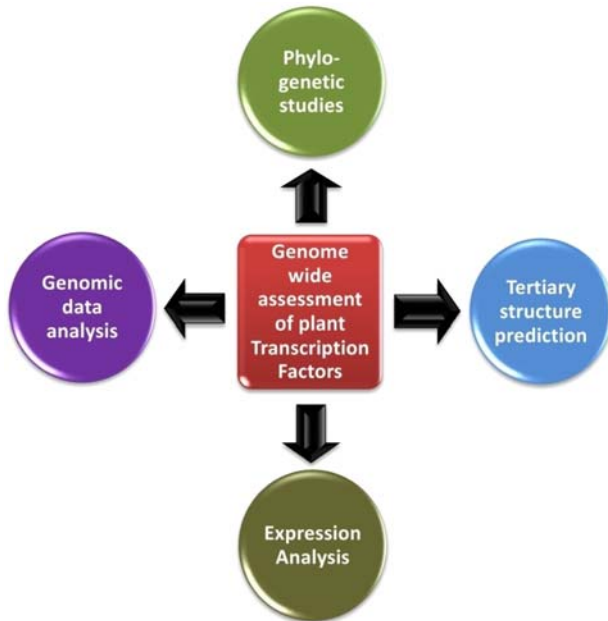
Sl. no.	Database name	Basic features	References
27.	DPTF	<ul style="list-style-type: none"> • Populus transcription factor. • Web: http://dptf.cbi.pku.cn 	Zhu, Guo, Gao, and Zhong (2007)
28.	STIFDB	<ul style="list-style-type: none"> • Stress-responsive Transcription factor database is a collection of genes and TFs responsive to several abiotic stresses. • Web: http://caps.ncbs.res.in/stifdb/brows.html • v2.0 includes some biotic stresses too. • Web: http://caps.ncbs.res.in/stifdb2/ 	Shameer, Ambika, Varghese, and Karaba (2009)
29.	Stress2TF	<ul style="list-style-type: none"> • Database for both biotic and abiotic stresses. • Web: http://csgenomics.ahau.edu.cn/Stress2TF 	Zhang, Yao, Fu, and Xuan (2018)
30.	iTAK	<ul style="list-style-type: none"> • This program recognizes transcription factors, transcriptional regulators, and protein kinases from nucleic acids and protein sequences. • In v1.6, rules have been updated for some TF families like HB and LIM. • This version fixes bugs in TFs/TRs classification to generate zip files. • v1.7 removed PCC classification. • Web: http://itak.feilab.net/cgi-bin/itak/index.cgi 	Zheng, Jiao, Sun, and Rosli (2016)
31.	AthTF	<ul style="list-style-type: none"> • This database offers predicted 3D models of TFs in Arabidopsis. • It possesses 2918 model structures having high confident score. • Web: http://sysbio.unl.edu/AthTF/ 	Lu, Yang, Yao, and Liu (2012)
32.	AthaMap	<ul style="list-style-type: none"> • It is whole-genome map of <i>Arabidopsis</i> TFs and small RNA-binding sites. • It represents a full list of 211 TFs with their references and screening results available on documentation page. • v7.0 (2012) includes micro-RNA target tool to identify miRNA in Arabidopsis genome. • Total identified TFBS in v7.0 is 1×10^7. • v8.0 (2016) is the latest with 5×10^7 TFBS of 207 different TFs and 32 TF families in database. • Web: http://www.athamap.de/ 	Steffens, Galuschka, Schindler, Bülow, and Hehl (2004)
33.	EXPath	<ul style="list-style-type: none"> • This portal maintains the expression data of model crops generated by microarray technique under different developmental and stress conditions. • In v2.0, the number of crops increased from three to six. • It provides tools for promoter analysis (PlantPAN) and compares expression profile by analyzing RNA-Seq and microarray data. • Correlation network construction within a gene group under various situations as well as information related to TFs of metabolic pathways also integrated into it. • Web: http://expath.itps.ncku.edu.tw/ 	Tseng, Li, Hung, and Chow (2020)
34.	FootprintDB	<ul style="list-style-type: none"> • It is the collection of 2422 TF sequences, 10,112 DNA-binding sites, and 3662 DNA motifs. • Web: http://floresta.eead.csic.es/footprintdbb 	Sebastian and Contreras-Moreira (2014)
35.	CGDB	<ul style="list-style-type: none"> • Coriander Genomics Database is the repository of genomic, transcriptomic, metabolomic, and functional data of coriander and carrot plant. • It includes 63 TF families of coriander and 61 of carrot. • Web: http://cgdb.bio2db.com/ 	Song, Nie, and Chen (2020)
36.	RARTF	<ul style="list-style-type: none"> • RIKEN Arabidopsis TF database. • Web: http://rarge.gsc.riken.jp/rartf/ 	Iida et al. (2005)
37.	DATF	<ul style="list-style-type: none"> • Database of Arabidopsis TF. 	Guo, He, Liu, and Bai (2005)

(Continued)

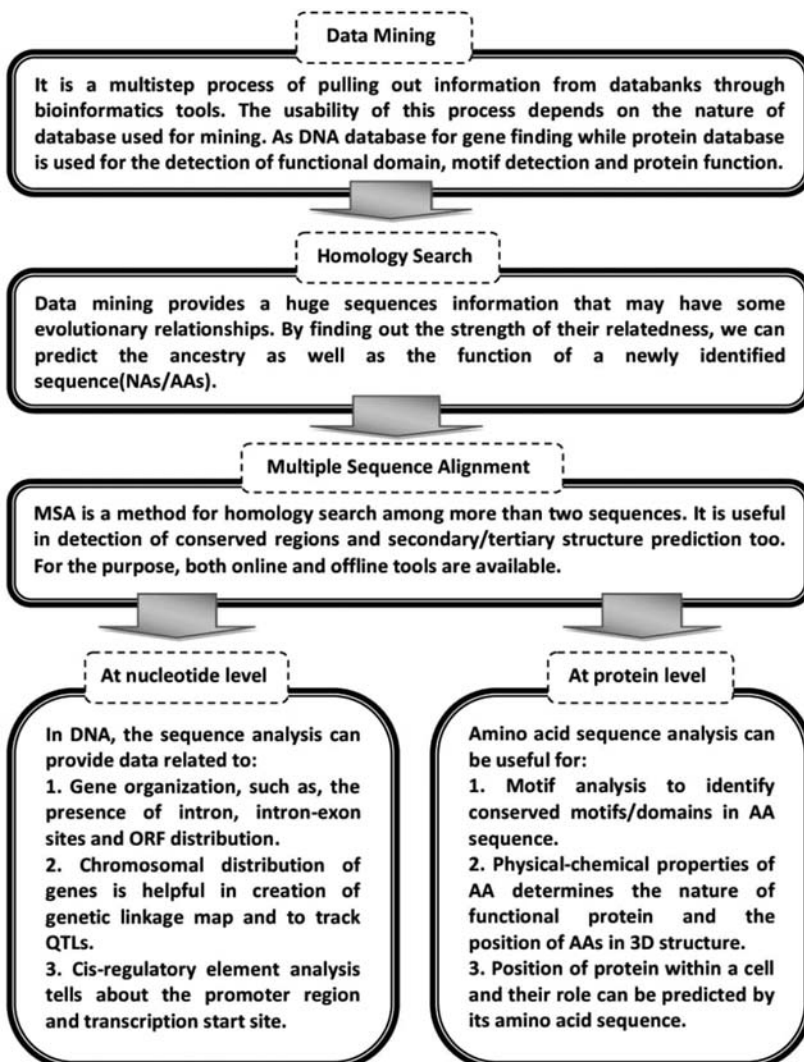
TABLE 6.3 (Continued)

Sl. no.	Database name	Basic features	References
		<ul style="list-style-type: none"> • Web: http://datf.cbi.pku.edu.cn/ 	
38.	TOBFAC	<ul style="list-style-type: none"> • Tobacco TF database contains a list of 2513 TFs classified in 64 gene families. • Web: http://compsysbio.achs.virginia.edu/tobfac/ 	Rushton, Bokowiec, Laudeman, and Brannock (2008)
39.	CnTDB	<ul style="list-style-type: none"> • Coconut Transcriptome Database has been developed to identify better traits against biotic stresses imposed by different pathogens. • It contains data of TF gene families, their expression, and role in different metabolic pathways. • Web: http://webtom.cabgrid.res.in/cntdb/ 	Verma, Jasrotia, Iquebal, and Jaiswal (2017)
40.	DBD	<ul style="list-style-type: none"> • DNA-binding domain database is helpful to predict sequence-specific TFs in available genomic sequences. • Web: http://www.transcriptionfactor.org/index.cgi?Home 	Kummerfeld and Teichmann (2006)
41.	realDB	<ul style="list-style-type: none"> • It is genome cum transcriptome database of red algae (Rhodophyceae). • Web: http://realdb.algaegenome.org/ 	Chen, Zhang, Chen, and Li (2018)
42.	YEASTRACT +	<ul style="list-style-type: none"> • Yeast Search Transcriptional Regulators And Consensus Tracking is a regulatory network database of <i>Saccharomyces cerevisiae</i>. • Web: http://www.yeasttract.com/ 	Monteiro, Oliveira, Pais, and Antunes (2020)
43.	RegulatorDB	<ul style="list-style-type: none"> • This database is a collection of tools to contemplate and explore expression profile of regulatory proteins in yeast mutants. • Web: http://wyrickbioinfo2.smb.wsu.edu/cgi-bin/RegulatorDB/cgi/home.pl 	Kemmeren, Sameith, van de Pasch, and Benschop (2014)
44.	YeTSFaSCo	<ul style="list-style-type: none"> • This database contains yeast TFs in position frequency matrix (PFM) or position weight matrix (PWM) formats. • Web: http://yetfasco.cabr.utoronto.ca/ 	Lee, Tillo, Bray, and Morse (2007)
45.	YeastSS	<ul style="list-style-type: none"> • It is Yeast Transcription Start Site database • Web: http://www.yeastss.org/ 	McMillan, Lu, Rodriguez, Ahn, and Lin (2019)
46.	ScerTF	<ul style="list-style-type: none"> • <i>Saccharomyces cerevisiae</i> Transcription Factor database is a collection of 196 TFs of yeast in PWMs format. • Web: http://stormo.wustl.edu/ScerTF/references/ 	Spivak and Stormo (2012)
47.	TRANSFAC	<ul style="list-style-type: none"> • It is Eukaryotic TF database having their DNA-binding site-related information. • Web: https://genexplain.com/transfac/ 	Wingender, Dietze, Karas, and Knüppel (1996)
48.	GTRD	<ul style="list-style-type: none"> • Gene Transcription Regulation Database. • Web: http://gtrd.biouml.org/# 	Kolmykov, Yevshin, Kulyashov, and Sharipov (2021)

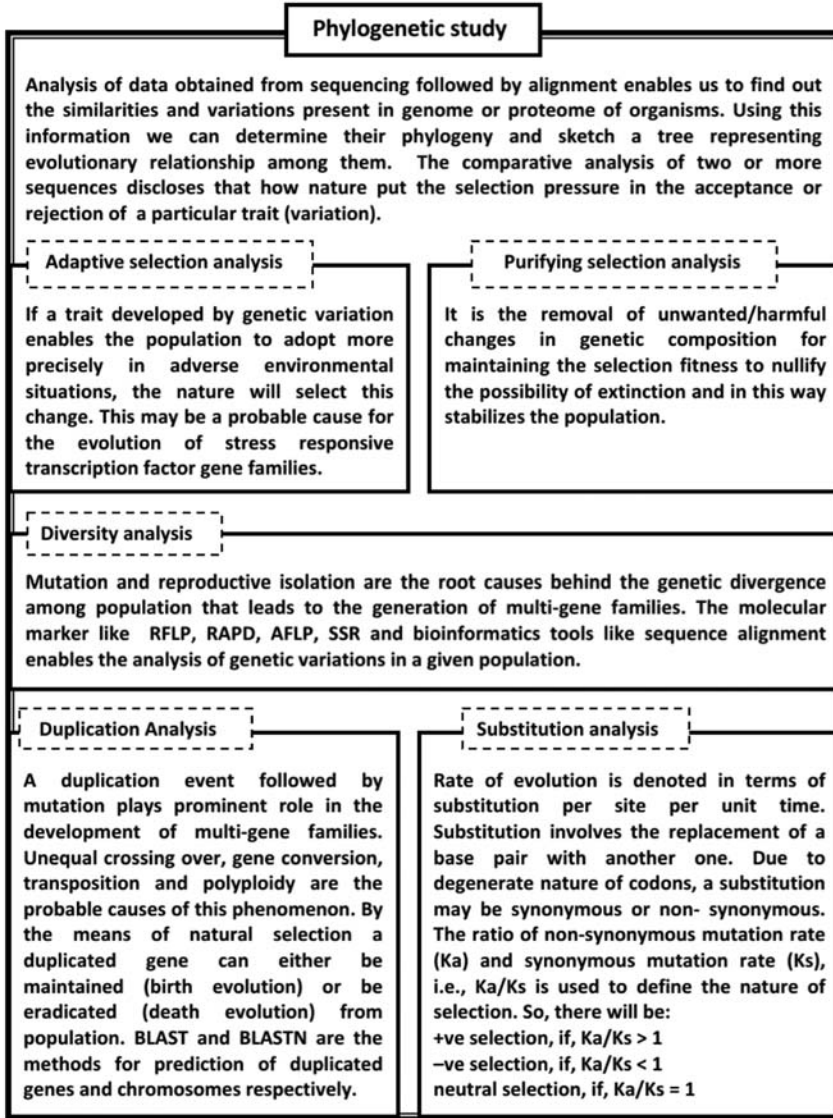
FIGURE 6.2 Genome-wide analysis of transcription factors.

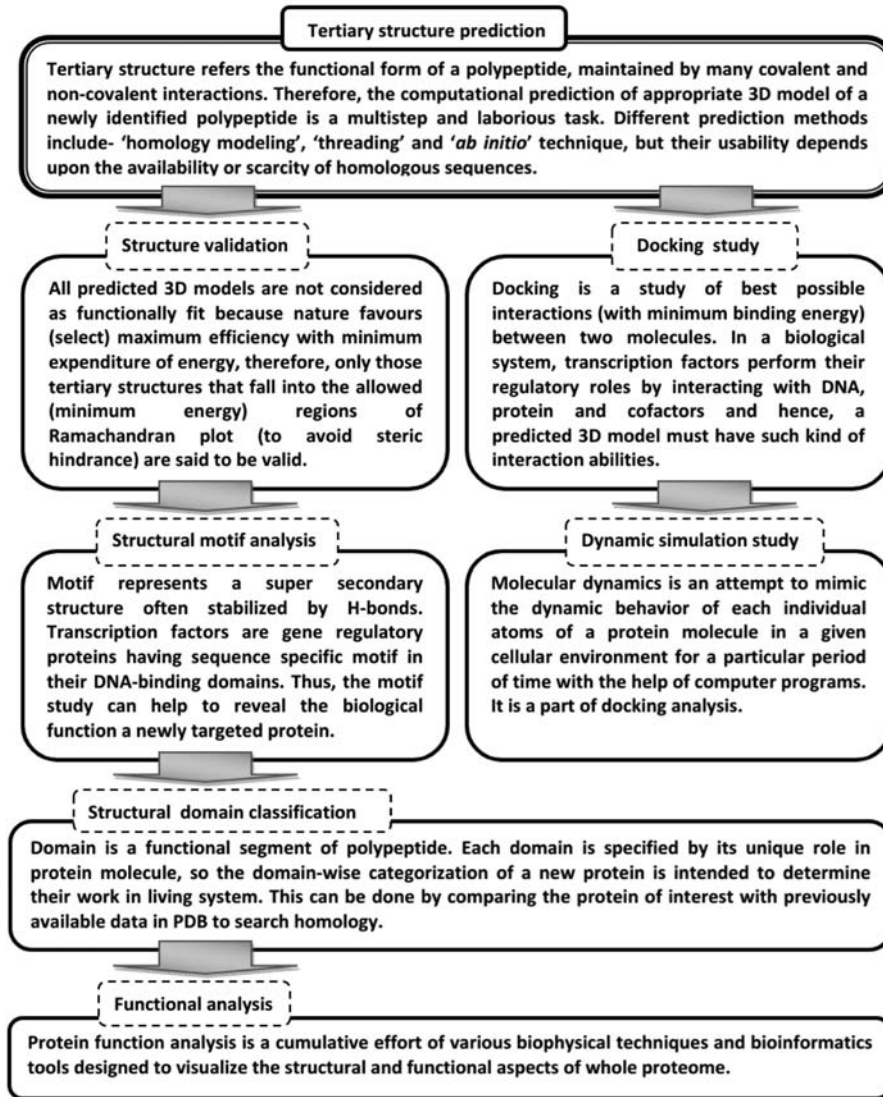


FLOW CHART 1 Genomic data analysis.



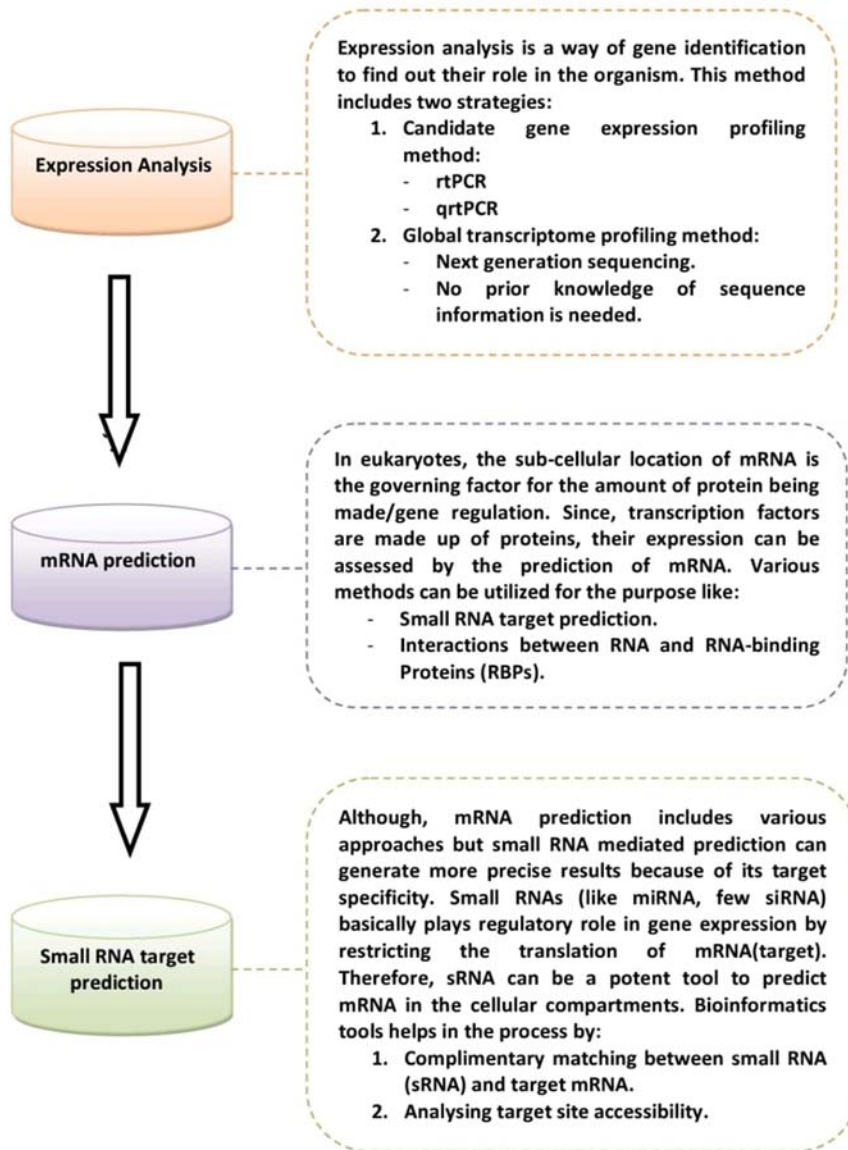
FLOW CHART 2 Phylogenetic studies.





FLOW CHART 3 Tertiary structure predictions.

FLOW CHART 4 Expression analysis.



The bioinformatics tools both online and offline, which are commonly used for the analysis of TF, are shown in Table 6.4.

6.5.1 Data mining by National Center for Biotechnology Information

Data mining includes sets of bioinformatics tools associated with deciphering important attributes from collected data which are generally sequences. There exist several databases varying in structure, design, and applications that could be processed and analyzed based on specific applications (Yang, Li, & Liu, 2020). The National Center for Biotechnology Information is one of the most commonly used platforms comprising a large collection of online resources exclusively for biological information and data. It includes GenBank and PubMed for the nucleic acid sequences and citations and abstracts published in life science journals, respectively. In this database, there is a provision for search and retrieval of desired information using Entrez system connected to around 34 different databases using E-utilities (Sayers, Beck, Bolton, & Bourexis, 2021).

TABLE 6.4 Some of the important bioinformatics tools used for the assessment of transcription factor gene families.

Sl. no.	Task	Mode	Bioinformatics tools
1.	Data mining	Online	NCBI http://www.ncbi.nlm.nih.gov/
2.	Homology search	Online	BLAST http://blast.ncbi.nlm.nih.gov/Blast.cgi
3.	Multiple sequence alignment	Online	Clustal http://www.ebi.ac.uk/Tools/msa/clustalo/
			Muscle http://www.ebi.ac.uk/Tools/msa/muscle/
			Kalign http://www.ebi.ac.uk/Tools/msa/kalign/
			Mafft http://www.ebi.ac.uk/Tools/msa/mafft/
			T-Coffee http://www.ebi.ac.uk/Tools/msa/tcoffee/
		Offline	Clustal
4.	Protein domain functional analysis	Online	Motif Scan http://myhits.isb-sib.ch/cgi-bin/motifscan
			InterPro http://www.ebi.ac.uk/interpro/
			Pfam http://pfam.sanger.ac.uk/search
			Prosite http://prosite.expasy.org/
			ScanProsite http://prosite.expasy.org/scanprosite/
			Smart http://smart.embl-heidelberg.de/
5.	Conserved domain search	Online	CD search http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
6.	Physical and chemical properties search	Online	ProtParam http://web.expasy.org/protparam/
7.	Motif search	Online	MEME http://meme-suite.org/tools/meme
			MAST http://meme-suite.org/tools/mast
			Tomtom http://meme-suite.org/tools/tomtom
8.	Ancestral sequence construction	Online	FASTML http://fastml.tau.ac.il/
9.	Adaptive and purifying selection analysis	Online	DATAMONKEY http://www.datamonkey.org/
10.	Gene structure display	Online	GSDS http://gsds.cbi.pku.edu.cn/
11.	Physical properties and representation server	Online	ProtScale http://web.expasy.org/protscale/

(Continued)

TABLE 6.4 (Continued)

Sl. no.	Task	Mode	Bioinformatics tools
12.	Location of signal peptide	Online	SignalP http://www.cbs.dtu.dk/services/SignalP/
13.	Subcellular location	Online	TargetP http://www.cbs.dtu.dk/services/TargetP/
14.	3D structure prediction	Online	I-TASSER http://zhanglab.ccmb.med.umich.edu/I-TASSER/
		Offline	MODELLER
15.	Ab initio protein folding and protein structure prediction	Online	QUARK http://zhanglab.ccmb.med.umich.edu/QUARK/
16.	Protein structure prediction	Online	LOMETS http://zhanglab.ccmb.med.umich.edu/LOMETS/
17.	Protein–ligand-binding site prediction	Online	COACH http://zhanglab.ccmb.med.umich.edu/COACH/
18.	Structure-based function prediction	Online	CO-FACTOR http://zhanglab.ccmb.med.umich.edu/COFACTOR/
19.	Fragment-guided MD simulation	Online	FG-MD http://zhanglab.ccmb.med.umich.edu/FG-MD/
20.	Ramachandran plot analysis	Online	RAM PAGE http://mordred.bioc.cam.ac.uk/~rapper/rampage.php
21.	Ramachandran plot and secondary structure analysis	Online	PDBsum https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html
22.	Homology modeling	Online	SWISS-MODEL http://swissmodel.expasy.org/
23.	Sterio-chemical quality of 3D structure	Online	PRO-CHECK http://services.mbi.ucla.edu/PROCHECK/
24.	Protein structure verification by crystallography	Online	ERRAT http://services.mbi.ucla.edu/ERRAT/
25.	Verification of 3D structure	Online	VERIFY 3D http://services.mbi.ucla.edu/Verify_3D/
26.	3D structure visualization	Offline	Discovery studio
27.	3D structure minimization	Offline	Chimera
28.	Gene Ontology annotation	Online	AgBase http://www.agbase.msstate.edu/
29.	Secondary structure prediction	Online	PSIPRED http://bioinf.cs.ucl.ac.uk/psipred/
			APSSP http://www.imtech.res.in/raghava/apssp2/
			Jpred http://www.compbio.dundee.ac.uk/jpred/
			SOPMA https://npsaprabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html
30.	Protein domain classification	Online	CATH http://www.cathdb.info/

(Continued)

TABLE 6.4 (Continued)

Sl. no.	Task	Mode	Bioinformatics tools
			Gene3D http://gene3d.biochem.ucl.ac.uk/Gene3D/
31.	Protein structural classification	Online	SCOP http://clavius.bc.edu/~clotelab/DiANNA/
32.	Cysteine state and disulfide bond partner prediction	Online	DiANNA http://clavius.bc.edu/~clotelab/DiANNA/
33.	Plant small RNA target analysis	Online	PsRNA Target http://plantgrn.noble.org/psRNATarget/
34.	miRNAs prediction	Online	MiREval http://mimima.centenary.org.au/mireval/
35.	Docking	Online	Molecular docking http://www.dockingserver.com/web
		Offline	AutoDock
36.	Cis-regulatory element analysis	Online	PLACE http://www.dna.affrc.go.jp/PLACE/
		Online	PlantCARE http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
37.	Gene prediction	Online	SOFTBERRY http://www.softberry.com/
38.	Substitution analysis	Offline	Ka–Ks calculator
			K-Estimator
			DnaSP

6.5.2 BLAST tool

BLAST (basic local alignment search tool) is an important bioinformatics tool commonly used for identifying the similarity among different sequences through a web interface or by means of an independent tool for comparing a user's query to the existing database of sequences (Altschul, Madden, Schaffer, & Zhang, 1997). Comparisons between pairs of sequences, and searches for regions of local similarity can be performed by BLAST (Pertsemlidis & Fondon III, 2001). By performing sequence similarity searches, identification of “homologous” proteins or genes can be achieved and common ancestry can be predicted. Homology among sequences gives an idea that sequences may be related by divergence from a common ancestor, or it may share common functional characteristics. Sequence homology searches serve as a key computational tool of molecular biology. They are very important, as their products and their high scoring alignments are used in a broad range of areas, from the assessment of evolutionary histories to the prediction of functions of genes and proteins, to the identification of possible drug targets (Bailey & Gribskov, 1998; Bayat, 2002; Pearson, 2013).

6.5.3 Multiple sequence alignment

Multiple sequence alignment identifies the presence of specific patterns or motifs based on the comparison of sequences revealing homology between new sequences and existing families of sequences. In most of the cases the query sequences are the proteins that are assumed to have evolutionary relationship based on conserved regions and share a common lineage that originates from the common ancestor. There is also provision of prediction of secondary and tertiary structures based in alignment for molecular evolutionary studies. Some of the tools used are T-Coffee, MAFFT, and ClustalW.

6.5.3.1 MAFFT

MAFFT is known to be highly accurate and possesses good-quality algorithm for multiple sequence alignment. It uses two innovative techniques, the fast Fourier transform (FFT) and fundamental scoring system. The FFT identifies homologous regions, while fundamental scoring system reduces time taken by CPU and increases the accuracy of alignments. It uses two-cycle heuristics, including the progressive method, that is, FFT-NS-2, and iterative refinement method, that is, FFT-NS-I (Kato, Misawa, Kuma, & Miyata, 2002).

6.5.3.2 T-Coffee

T-Coffee stands for tree-based consistency objective function for alignment evolution. It uses an iterative multiple sequence alignment algorithm involving data sources from both global and local pairwise alignments. It is comparatively better than ClustalW with reference to the level of accuracy though has disadvantage of weak scalability. It can align a maximum of 100 sequences (Daugelaite, O' Driscoll, & Sleator, 2013).

6.5.3.3 Clustal

The Clustal programs like ClustalW can perform automatic multiple alignment of sets of nucleotide or amino acid sequences using a user-friendly simple text menu system portable to almost all computer systems (Thompson, Higgins, & Gibson, 1994). In Clustal X a graphical user interface is used to perform multiple alignments. Clustal W and Clustal X have been developed with the provision of similar version-numbering system for synchronizing changes like bug fixing, improvements, and additions (Thompson, Gibson, Plewniak, Jeanmougin, & Higgins, 1997). Clustal Omega aligns two profile Hidden Markov Models (HMMs), instead of a profile–profile comparison and has improved sensitivity and alignment quality (Soding, 2005).

6.5.3.4 MUSCLE

MUSCLE stands for Multiple Sequence Comparison by Log-Expectation. It operates by using two distance measures, kmer distance and Kimura distance, exclusively for unaligned and aligned pairs of sequences, respectively. Further it involves the preparation of guide trees by UPGMA (unweighted pair group method with arithmetic mean) method. Kimura distance method is preferred as compared to kmer (Edgar, 2004).

6.5.3.5 Kalign

Kalign algorithm is used for performing multiple sequence alignment using standard progressive methods such as pairwise distances. It uses k-tuple method for calculations as adopted from ClustalW and involves construction of guide tree using either neighbor-joining method or UPGMA. It uses Wu–Manber approximate string-matching algorithm is a unique feature (Daugelaite et al., 2013).

6.5.4 Physicochemical properties analysis

Based on the protein sequences, in silico prediction of several physicochemical features like molecular weight, pI, total number of negative charged residues, total number of positively charged residues, extinction coefficient, instability index, aliphatic index, and GRAVY (Grand average of hydropathy) can be determined. The ExPASy (Expert Protein Analysis System) is a web server that provides access to a variety of databases and analytical tools useful for proteins and proteomics. It includes SWISSPROT and TrEMBL, SWISS-2D PAGE, PROSITE, ENZYME, and the SWISS-MODEL repository. Some of the analytical tools commonly used are Compute pI/MW, ProtParam, PeptideMass, PeptideCutter, ProtScale, etc. (Gasteiger et al., 2005).

6.5.5 Motif and domain prediction

The functional identity of the protein is based on short conserved sequences referred to as motifs and domains. Domains are typically longer than motifs. Motifs are often associated with a distinct structural site performing a particular function as in the case of TFs like Zn finger motif having 10–20 amino acids. A domain also reflects a comparatively larger conserved sequence pattern with a distinct functional and structural unit. TF prediction from genomes sequences is based on conserved domain sequences of respective TFs.

6.5.5.1 *InterPro*

InterPro is a web-based tool used for the identification of protein domains and various functional sites using several databases like PROSITE, PRINTS, ProDom, Pfam, and SMART (Simple Modular Architecture Research Tool). The pattern matching is accomplished by the combination of expressions, fingerprints, profiles, and HMMs. An InterPro search provides a graphical output summarizing motif matches (Biswas, O'Rourke, Camon, & Fraser, 2002).

6.5.5.2 *SMART*

SMART comprises HMM profiles constructed from manually refined protein domain alignments. Here, alignments are mainly constructed based on tertiary structures or based on PSI (position specific iterative)-BLAST profiles. The SMART database consists of several independent collections of HMMs with greater emphasis on signaling, extracellular, and chromatin-associated motifs and domains. Graphical representation of domains is the output of sequence search (Xiong, 2006).

6.5.5.3 *MEME Suite*

The MEME Suite web server serves as an integrated portal for elucidating motifs and domains. It can reveal sequence motifs with DNA-binding sites and also protein interaction domains. The upgraded version involves GLAM2 algorithm that allows the discovery of motifs containing gaps. It uses three different sequence scanning algorithms, MAST (motif alignment and search tool), FIMO (find individual motif occurrence), and GLAM2SCAN (scanning with gapped motifs). It is frequently used for the analysis of TF motifs, and further functional elucidation can be achieved by Gene Ontology (GO) terms using the motif-GO term association tool GOMO (gene ontology for motifs). The output of this web server can be represented by means of sequence LOGOS for each discovered motif.

6.5.6 **In silico structure prediction of proteins**

The knowledge of 3D structure of a protein provides an insight into the function of the protein, and there exist both experimental methods like X-ray crystallography, NMR and in silico based methods like homology modeling, threading, and ab initio modeling.

6.5.6.1 *I-TASSER*

I-TASSER stands for the iterative threading assembly refinement server used for automated protein structure and functions prediction. It is based on the sequence-to-structure-to-function pattern determination. It first generates 3D atomic models utilizing multiple threading alignments and iterative structural assembly simulations using amino acid sequences. The function of the protein is then determined by structurally comparing the 3D models with other available known proteins. The output is represented by full-length secondary and tertiary structure predictions, functional annotations on ligand-binding sites, GO terms, and enzyme commission numbers (Roy, Kucukural, & Zhang, 2010).

6.5.6.2 *Modeller*

In the case of lack of experimentally determined structure, comparative or homology modeling is an important tool for the determination of 3D model for a protein (Misura, Chivian, Rohl, Kim, & Baker, 2006). It predicts the 3D structure of a given target protein sequence on the basis of its alignment to one or more proteins of known structure. It comprises four steps, fold assignment revealing similarity between target and known template structure, alignment, building model, and predicting model errors (Marti-Renom, Stuart, Fiser, & Sanchez, 2000).

6.5.6.3 *PDBsum*

It is a web server revealing image-based structural information like protein secondary structure, protein–ligand and protein–protein interactions about the entries of Protein Data Bank. It includes a complete PROCHECK assessment of each protein's geometry (Laskowski, MacArthur, Moss, & Thornton, 1993). The Ramachandran plot in PDBsum can validate the predicted 3D structure and can be explored interactively in RasMol, PyMOL, and a JavaScript viewer called 3Dmol.js. (Laskowski, Jabłońska, Pravda, Vařeková, & Thornton, 2018).

6.5.7 Gene predictions

Analyzing genome sequences to fish out genes needs computational methods. Bioinformatics tools are available for the prediction of genes from the sequences provided by identifying some of the essential attributes of genes. This tool should provide information about the protein coding regions along with several functional sites. Computational gene prediction methods are based either on sequence similarity searches or gene structure and signal-based searches popularly referred to as *ab initio* gene finding. Sequence similarity search-based gene prediction can provide an insight into similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome. It relies on the fact that those functional exon regions are more conserved evolutionarily than nonfunctional intergenic or intron regions. The second method uses gene structure as a template to detect genes and depends on two types of sequence information, signal and content sensors. Signal sensors basically refer to short sequence motifs like splice sites, branch points, polypyrimidine tracts, start codons, and stop codons. Content sensors are generally used for exon detection and have the provision for separating coding sequences from the surrounding noncoding sequences using appropriate statistical detection algorithms (Wang, Chen, & Li, 2004; Wang, Hong, & Han, 2004).

6.5.8 Gene duplication and functional divergence studies

The nonsynonymous (K_a) and synonymous (K_s) substitution rates provide an insight into evolutionary dynamics of protein-coding sequences across closely related and yet diverged species (Fay & Wu, 2003; Kimura, 1983; Li, He, Wang, & Wang, 2013). Based on values of K_a and K_s and their ratio (K_a/K_s), information about neutral mutation ($K_a = K_s$), negative (purifying) selection (K_a less than K_s), and positive (diversifying) selection (K_a exceeds K_s) can be obtained. K_a/K_s Calculator is software that calculates nonsynonymous (K_a) and synonymous (K_s) substitution rates through model selection and model averaging (Zhang et al., 2006). Gene duplication events in the protein family can be tested by type I functional divergence through Diverge version 2.0 software. Three methods such as single likelihood ancestor counting, fixed-effect likelihood, and random-effect likelihood are generally employed to select individual codons, using the default settings of the DataMonkey web-based server (Delpont, Poon, Frost, & Kosakovsky Pond, 2010).

6.6 Conclusion

TFs are known to be an important element associated with gene regulation by interacting with specific sequences of promoters of the concerned genes. The importance of TF in gene expression and regulation was realized based on the fact that 5%–10% of whole-genome sequence represents genes coding for TFs. There are several types of TFs, some are common in plants and animals, while there exist plant-specific TFs known to be associated with functions influencing growth and development of plants. The presence of specific type of DBD in the TFs is considered to be one important criterion for classification of TFs. Bioinformatics-based assessment of these plant-specific TFs gained momentum with the development of crop-specific TF databases, freely available to the researchers. The reports of genome-wide *in silico* prediction, bioinformatics-based sequence characterization, wet lab-based cloning, and expression profiling of several plant-specific TFs, representing crops, genome sequences of which have been deciphered, are substantially increasing. The potentials of stress-responsive TFs in developing biotic- or abiotic-tolerant crops by transgenic approach have been realized by the plant biotechnologists and are being investigated extensively.

References

- Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2003). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *The Plant Cell*, 15(1), 63–78.
- Agarwal, P., Baranwal, V. K., & Khurana, P. (2019). Genome-wide analysis of bZIP transcription factors in wheat and functional characterization of TabZIP under abiotic stress. *Scientific Reports*, 9, 4608.
- Agarwal, P. K., Gupta, K., Lopato, S., & Agarwal, P. (2017). Dehydration responsive element binding transcription factors and their applications for the engineering of stress tolerance. *Journal of Experimental Botany*, 68(9), 2135–2148.
- Agarwal, P. K., Shukla, P. S., Gupta, K., & Jha, B. (2013). Bioengineering for salinity tolerance in plants: State of the art. *Molecular Biotechnology*, 54(1), 102–123.
- Ali, Z., Sarwat, S. S., Karim, I., Faridi, R., et al. (2016). Functions of plant's bZIP transcription factors. *Pakistan Journal of Agricultural Sciences*, 53, 303–314.

- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389–3402.
- Alves, M. S., Dadalto, S. P., Goncalves, A. B., De Souza, G. B., et al. (2013). Plant bZIP transcription factors responsive to pathogens: A review. *International Journal of Molecular Sciences*, 14, 7815–7828.
- Arora, R., Agarwal, P., Ray, S., Singh, A. K., et al. (2007). MADS-box gene family in rice: Genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics*, 8, 242.
- Arrey-Sales, O., Caris-Maldonado, J. C., Hernandez-Rojas, B., & Gonzales, E. (2021). Comprehensive genome-wide exploration of C2C2 zinc finger family in grapevine (*Vitis vinifera* L): Insights into the roles in the pollen development regulation. *Genes*, 12, 302.
- Bailey, T. L., & Gribskov, M. (1998). Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics (Oxford, England)*, 14(1), 48–54.
- Baldoni, E., Genga, A., & Cominelli, E. (2015). Plant MYB transcription factors: Their role in drought response mechanisms. *International Journal of Molecular Sciences*, 16(7), 15811–15851. Available from <https://doi.org/10.3390/ijms160715811>. Available from 26184177, PMCID: PMC4519927.
- Bastola, D. R., Pethe, V. V., & Winicov, I. (1998). Alfin1, a novel zinc-finger protein in alfalfa roots that binds to promoter elements in the salt-inducible MsPRP2 gene. *Plant Molecular Biology*, 38, 1123–1135.
- Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. In clinical review. *British Medical Journal*, 324, 1018–1022.
- Bhawna., Bonthala, V. S., & Gajula, M. P. (2016). PvTFDB: A *Phaseolus vulgaris* transcription factors database for expediting functional genomics in legumes. *Database (Oxford)*, 27.
- Biswas, M., O'Rourke, J. F., Camon, E., Fraser, G., et al. (2002). Applications of InterPro in protein annotation and genome analysis. *Briefings in Bioinformatics*, 3(3), 285–295.
- Cai, X., Zhang, Y., Zhang, C., Zhang, T., et al. (2013). Genome-wide analysis of plant-specific Dof transcription factor family in tomato. *Journal of Integrative Plant Biology*, 55(6), 552–566.
- Chen, C., Chen, X., Han, J., Lu, W., et al. (2020). Genome-wide analysis of the WRKY gene family in the cucumber genome and transcriptome-wide identification of the WRKY transcription factors that respond to biotic and abiotic stresses. *BMC Plant Biology*, 20, 443.
- Chen, F., Zhang, J., Chen, J., Li, X., et al. (2018). realDB: A genome and transcriptome resource for the red algae (phylum Rhodophyta). *Database (Oxford)*, 1.
- Chen, L., Song, Y., Li, S., Zhang, L., Zou, C., & Yu, D. (2012). The role of WRKY transcription factors in plant abiotic stresses. *Biochimica et Biophysica Acta (BBA)*, 1819(2), 120–128.
- Chen, P., & Liu, Q. Z. (2019). Genome-wide characterization of the WRKY gene family in cultivated strawberry (*Fragaria x ananassa* Duch.) and the importance of several group III members in continuous cropping. *Scientific Reports*, 9, 8423.
- Chen, Q., Wang, Q., Xiong, L., & Lou, Z. (2011). A structural view of the conserved domain of rice stress-responsive NAC1. *Protein and Cell*, 2(1), 55–63.
- Chen, W., Chen, Z., Luo, F., Liao, M., et al. (2019). RicetissueTFDB: A genome-wide identification of tissue-specific transcription factors in rice. *The Plant Genome*, 12(1).
- Chu, Y., Xiao, S., Su, H., Liao, B., et al. (2018). Genome-wide characterization and analysis of bHLH transcription factors in *Panax ginseng*. *Acta Pharmaceutica Sinica B (APSB)*, 8(4), 666–677.
- Ciftci-Yilmaz, S., Morsy, M. R., Song, L., Coutu, A., et al. (2007). The EAR-motif of the Cys2/His2-type zinc finger protein Zat7 plays a key role in the defense response of Arabidopsis to salinity stress. *Journal of Biological Chemistry*, 282, 9260–9268.
- da Silva, D. C., da Silveira Falavigna, V., Fasoli, M., Buffon, V., et al. (2016). Transcriptome analyses of the Dof-like gene family in grapevine reveal its involvement in berry, flower and seed development. *Horticulture Research*, 3, 16042.
- Daugelaite, J., O'Driscoll, A., & Sleator, R. D. (2013). An overview of multiple sequence alignments and cloud computing in bioinformatics. *International Scholarly Research Notices*. Available from <https://doi.org/10.1155/2013/615630>, Article ID 615630, 14 pages.
- Davletova, S., Schlauch, K., Coutu, J., & Mittler, R. (2005). The zinc-finger protein Zat12 plays a central role in reactive oxygen and abiotic stress signaling in Arabidopsis. *Plant Physiology*, 139, 847–856.
- De Paolis, A., Caretto, S., Quarta, A., Di Sansebastiano, G. P., et al. (2020). Genome-wide identification of WRKY genes in *Artemisia annua*: Characterization of a putative ortholog of AtWRKY40. *Plants*, 9, 1669.
- Delpont, W., Poon, A. F. Y., Frost, S. D. W., & Kosakovsky Pond, S. L. (2010). Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics (Oxford, England)*, 26(19), 2455–2457.
- Diao, W., Snyder, J. C., Wang, S., Liu, J., et al. (2018). Genome-wide analyses of the NAC transcription factor gene family in pepper (*Capsicum annuum* L.): Chromosome location, phylogeny, structure, expression patterns, Cis-elements in the promoter, and interaction network. *International Journal of Molecular Sciences*, 19(4), 1028.
- Dong, N., Liu, X., Lu, Y., Du, L., et al. (2010). Overexpression of TaPIEP1, a pathogen-induced ERF gene of wheat, confers host-enhanced resistance to fungal pathogen *Bipolaris sorokiniana*. *Functional & Integrative Genomics*, 10(2), 215–226.
- Duan, M. R., Nan, J., Liang, Y. H., Mao, P., et al. (2007). DNA binding mechanism revealed by high resolution crystal structure of *Arabidopsis thaliana* WRKY1 protein. *Nucleic Acids Research*, 35(4), 1145–1154.
- Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., et al. (2010). MYB transcription factors in Arabidopsis. *Trends in Plant Science*, 15(10), 573–581.
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797.

- Faraji, S., Filiz, E., Kazemitabar, S. K., Vannozi, A., et al. (2020). The AP2/ERF gene family in *Triticum durum*: Genome-wide identification and expression analysis under drought and salinity stresses. *Genes*, *11*, 1464.
- Fay, J. C., & Wu, C. I. (2003). Sequence divergence, functional constraint, and selection in protein evolution. *Annual Review of Genomics and Human Genetics*, *4*, 213–235.
- Feller, A., Machemer, K., Braun, E. L., & Grotewold, E. (2011). Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. *The Plant Journal*, *66*(1), 94–116.
- Feng, K., Hou, X. L., Xing, G. M., & Liu, J. X. (2020). Advances in AP2/ERF super-family transcription factors in plant. *Critical Reviews in Biotechnology*, *40*(6), 750–776.
- Fernández-Calvo, P., Chini, A., Fernández-Barbero, G., et al. (2011). The Arabidopsis bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. *The Plant Cell*, *23*, 701–715.
- Fornes, O., Castro-Mondragon, J. A., Khan, A., et al. (2019). JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *47*(D1). Available from <https://doi.org/10.1093/nar/gkz1001>.
- Gao, G., Zhong, Y., Guo, A., Zhu, Q., et al. (2006). DRTF: A database of rice transcription factors. *Bioinformatics (Oxford, England)*, *22*(10), 1286–1287.
- Gasteiger, E., et al. (2005). Protein identification and analysis tools on the ExPASy server. In J. M. Walker (Ed.), *The proteomics protocols handbook*. Springer protocols handbooks. Humana Press.
- Gayali, S., Acharya, S., Lande, N. V., et al. (2016). CicerTransDB 1.0: A resource for expression and functional study of chickpea transcription factors. *BMC Plant Biology*, *16*, 169.
- Gomez-Cano, F., Carey, L., Lucas, K., García Navarrete, T., et al. (2020). CamRegBase: A gene regulation database for the biofuel crop, *Camelina sativa*. *Database (Oxford)*, *11*, baaa075.
- Guerin, C., Roche, J., Allard, V., Ravel, C., et al. (2019). Genome-wide analysis, expansion and expression of the NAC family under drought and heat stresses in bread wheat (*T. aestivum* L.). *PLoS One*, *14*(3), e0213390.
- Guignon, V., Toure, A., Droc, G., Dufayard, J. F., et al. (2020). GreenPhylDB v5: A comparative pangenomic database for plant genomes. *Nucleic Acids Research*, *49*, D1464–D1471.
- Guo, A., He, K., Liu, D., Bai, S., et al. (2005). DATF: A database of Arabidopsis transcription factors. *Bioinformatics (Oxford, England)*, *21*(10), 2568–2569.
- Guo, X. J., & Wang, J. R. (2017). Global identification, structural analysis and expression characterization of bHLH transcription factors in wheat. *BMC Plant Biology*, *17*, 90.
- Gupta, S., Arya, G. C., Malviya, N., Bisht, N. C., & Yadav, D. (2016). Molecular cloning and expression profiling of multiple *Dof* genes of *Sorghum bicolor* (L) Moench. *Molecular Biology Reports*, *43*(8), 767–774.
- Gupta, S., Kushwaha, H., Singh, V. K., Bisht, N. C., et al. (2014). Genome wide in silico characterization of *Dof* transcription factor gene family of sugarcane and its comparative phylogenetic analysis with Arabidopsis, rice and sorghum. *Sugar Tech*, *16*(4), 372–384.
- Gupta, S., Malviya, N., Kushwaha, H., Nasim, J., et al. (2015). Insights into structural and functional diversity of *Dof* (DNA binding with one finger) transcription factor. *Planta*, *241*(3), 549–562.
- Gupta, S., Mishra, V. K., Kumari, S., Raavi, et al. (2019). Deciphering genome-wide WRKY gene family of *Triticum aestivum* L. and their functional role in response to abiotic stress. *Journal of Genetics and Genomics*, *41*, 79–94.
- He, X., Li, J. J., Chen, Y., Yang, J. Q., et al. (2019). Genome-wide analysis of the WRKY gene family and its response to abiotic stress in buckwheat (*Fagopyrum tataricum*). *Open Life Sciences*, *14*, 80–96.
- Hernando-Amado, S., González-Calle, V., Carbonero, P., & Barrero-Sicilia, C. (2012). The family of DOF transcription factors in *Brachypodium distachyon*: Phylogenetic comparison with rice and barley DOFs and expression profiling. *BMC Plant Biology*, *12*, 202.
- Huang, L., Wang, Y., Wang, W., Zhao, X., et al. (2018). Characterization of transcription factor gene OsDRAP1 conferring drought tolerance in rice. *Frontiers of Plant Science*, *9*, 94.
- Iida, K., Seki, M., Sakurai, T., Satou, M., et al. (2005). RARTF: Database and tools for complete sets of Arabidopsis transcription factors. *DNA Research: an International Journal for Rapid Publication of Reports on Genes and Genomes*, *12*(4), 247–256.
- Jain, D., Roy, N., & Chattopadhyay, D. (2009). CaZF, a plant transcription factor functions through and parallel to HOG and calcineurin pathways in *Saccharomyces cerevisiae* to provide osmotolerance. *PLoS One*, *4*, e5154.
- Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., et al. (2002). bZIP transcription factors in Arabidopsis. *Trends in Plant Science*, *7*(3), 106–111.
- Jia, J., Tong, C., Wang, B., et al. (2004). Hedgehog signalling activity of Smoothened requires phosphorylation by protein kinase A and casein kinase I. *Nature*, *432*, 1045–1050.
- Jiang, Y., Zeng, B., Zhao, H., Zhang, M., et al. (2012). Genome-wide transcription factor gene prediction and their expressional tissue-specificities in maize. *Journal of Integrative Plant Biology*, *54*, 616–630.
- Jiao, Z., Wang, L., Du, H., & Wang, Y. (2020). Genome-wide study of C2H2 zinc finger gene family in *Medicago truncatula*. *BMC Plant Biology*, *20*, 401.
- Jin, J. P., Tian, F., Yang, D. C., Meng, Y. Q., et al. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, *45*(D1), D1040–D1045.
- Kappel, S., Eggeling, R., Rimpler, F., Groth, M., et al. (2021). DNA-binding properties of the MADS-domain transcription factor SEPALLATA3 and mutant variants characterized by SELEX-seq. *Plant Molecular Biology*, *105*, 543–557.

- Katiyar, A., Smita, S., Lenka, S. K., Rajwanshi, R., et al. (2012). Genome-wide classification and expression analysis of MYB transcription factor families in rice and Arabidopsis. *BMC Genomics*, *13*, 544.
- Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.
- Kemmeren, P., Sameith, K., van de Pasch, L. A., Benschop, J. J., et al. (2014). Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, *157*(3), 740–752.
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge UK: Cambridge University Press.
- Kolchanov, N. A., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., et al. (2002). Transcription Regulatory Regions Database (TRRD): Its status in 2002. *Nucleic Acids Research*, *30*(1), 312–317.
- Kolmykov, S., Yevshin, I., Kulyashov, M., Sharipov, R., et al. (2021). GTRD: An integrated view of transcription regulation. *Nucleic Acids Research*, *49*(D1), D104–D111.
- Kummerfeld, S. K., & Teichmann, S. A. (2006). DBD: A transcription factor prediction database. *Nucleic Acids Research*, *34*, D74–D81.
- Kushwaha, H., Gupta, S., Singh, V. K., Rastogi, S., & Yadav, D. (2011). Genome wide identification of Dof transcription factor gene family in sorghum and its comparative phylogenetic analysis with rice and Arabidopsis. *Molecular Biology Reports*, *38*, 5037–5053.
- Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S., & Thornton, J. M. (2018). PDBsum: Structural summaries of PDB entries. *Protein Science*, *27*(1), 129–134.
- Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK – A program to check the stereochemical quality of protein structures. *The Journal of Applied Crystallography*, *26*, 283–291.
- Lee, W., Tillo, D., Bray, N., Morse, R. H., et al. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*, *39*(10), 1235–1244.
- Lescot, M., Hais, P. D., Thijs, G., Marchal, K., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Research*, *30*(1), 325–327.
- Li, B., Fan, R., Yang, Q., Hu, C., et al. (2020). Genome-wide identification and characterization of the NAC transcription factor family in *Musa acuminata* and expression analysis during fruit ripening. *International Journal of Molecular Sciences*, *21*, 634.
- Li, J., Xiong, Y., Li, Y., Ye, S., et al. (2019). Comprehensive analysis and functional studies of WRKY transcription factors in *Nelumbo nucifera*. *International Journal of Molecular Sciences*, *20*, 5006.
- Li, P., Chai, Z., Lin, P., Huang, C., et al. (2020). Genome-wide identification and expression analysis of AP2/ERF transcription factors in sugarcane (*Saccharum spontaneum* L.). *BMC Genomics*, *21*, 685.
- Li, W. T., He, M., Wang, J., & Wang, Y. P. (2013). Zinc finger protein (ZFP) in plants—A review. *Plant Omics*, *6*(6).
- Li, X., Duan, X., Jiang, H., Sun, Y., et al. (2006). Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. *Plant Physiology*, *141*, 1167–1184.
- Liao, Y., Zou, H., Wei, W., Hao, Y. J., et al. (2008). Soybean GmbZIP44, GmbZIP62 and GmbZIP78 genes function as negative regulator of ABA signaling and confer salt and freezing tolerance in transgenic Arabidopsis. *Planta*, *228*, 225–240.
- Lin, Z., Cao, D., Damaris, R. N., & Yang, P. (2020). Genome-wide identification of MADS-box gene family in sacred lotus (*Nelumbo nucifera*) identifies a SEPALLATA homolog gene involved in floral development. *BMC Plant Biology*, *20*, 497.
- Ling, L., Song, L., Wang, Y., & Guo, C. (2017). Genome-wide analysis and expression patterns of the NAC transcription factor family in *Medicago truncatula*. *Physiology and Molecular Biology of Plants*, *23*, 343–356.
- Liu, M., Ma, Z., Sun, W., Huang, L., et al. (2019). Genome-wide analysis of the NAC transcription factor family in Tartary buckwheat (*Fagopyrum tataricum*). *BMC Genomics*, *20*, 113.
- Liu, M., Sun, W., Ma, Z., Zheng, T., et al. (2019). Genome-wide investigation of the AP2/ERF gene family in Tartary buckwheat *Fagopyrum tataricum*. *BMC Plant Biology*, *19*, 84.
- Liu, Y., Liu, D., Hu, R., Hua, C., et al. (2017). AtGIS, a C2H2 zinc-finger transcription factor from Arabidopsis regulates glandular trichome development through GA signaling in tobacco. *Biochemical and Biophysical Research Communications*, *483*(1), 209–215.
- Lu, B., Wang, Y., Zhang, G., Feng, Y., et al. (2020). Genome-wide identification and expression analysis of the strawberry FvbZIP gene family and the role of key gene FabZIP46 in fruit resistance to gray mold. *Plants*, *9*, 1199.
- Lu, T., Yang, Y., Yao, B., Liu, S., et al. (2012). Template-based structure prediction and classification of transcription factors in *Arabidopsis thaliana*. *Protein Science*, *21*(6), 828–838.
- Luan, Q., Chen, C., Liu, M., Li, Q., et al. (2019). CsWRKY50 mediates defense responses to *Pseudoperonospora cubensis* infection in *Cucumis sativus*. *Plant Science (Shannon, Ireland)*, *279*, 59–69.
- Ma, J., Li, M. Y., Wang, F., Tang, J., & Xiong, A. S. (2015). Genome-wide analysis of Dof family transcription factors and their responses to abiotic stresses in Chinese cabbage. *BMC Genomics*, *16*, 33.
- Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., et al. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, *29*, 291–325.
- McMillan, J., Lu, Z., Rodriguez, J. S., Ahn, T., & Lin, Z. (2019). YeastTSS: An integrative web database of yeast transcription start sites. *Database (Oxford)*.
- Miller, J., McLachlan, A. D., & Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *The EMBO Journal*, *4*(6), 1609–1614.
- Min, X., Jin, X., Zhang, Z., & Wei, X. (2020). Genome-wide identification of NAC transcription factor family and functional analysis of the abiotic stress-responsive genes in *Medicago sativa* L. *Journal of Plant Growth Regulation*, *39*, 324–337.

- Misura, K. M., Chivian, D., Rohl, C. A., Kim, D. E., & Baker, D. (2006). Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 5361–5366.
- Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., et al. (2009). In silico analysis of transcription factor repertoire and prediction of stress responsive transcription factors in soybean. *DNA Research: an International Journal for Rapid Publication of Reports on Genes and Genomes*, 16(6), 353–369.
- Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., et al. (2010). Legume TFDB. *Bioinformatics (Oxford, England)*, 26, 290–291.
- Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., et al. (2011). Gramineae TFDB: An integrative database for functional genomics and comparative genomics of transcription factors from *Brachypodium distachyon*, maize, rice, sorghum, barley and wheat. *Plant and Cell Physiology*.
- Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., et al. (2013). TreeTFDB: An integrative database of the transcription factors from six economically important tree crops for functional predictions and comparative and functional genomics. *DNA Research*, 20(2), 151–162.
- Mohanta, T. K., Yadav, D., Khan, A., Hashem, A., et al. (2020). Genomics, molecular and evolutionary perspective of NAC transcription factors. *PLoS One*, 15(4), e0231425.
- Monteiro, P. T., Oliveira, J., Pais, P., Antunes, M., et al. (2020). YEASTRACT + : A portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Research*, 48(D1), D642–D649.
- Moreno-Risueno, M. A., Martinez, M., Vicente-Carbajosa, J., & Carbonero, P. (2007). The family of DOF transcription factors: from green unicellular algae to vascular plants. *Molecular Genetics and Genomics*, 277, 379–39.
- Muino, J. M., Smaczniak, C., Angenent, G. C., Kaufmann, K., & van Dijk, A. D. (2014). Structural determinants of DNA recognition by plant MADS-domain transcription factors. *Nucleic Acids Research*, 42(4), 2138–2146.
- Najafi, S., Sorkheh, K., & Nasernakhaei, F. (2018). Characterization of the APETALA2/ethylene responsive factor (AP2/ERF) transcription factor family in sunflower. *Scientific Reports*, 8, 11576.
- Nakashima, K., Tran, L.-S. P., Nguyen, D. V., Fujita, M., et al. (2007). Functional analysis of a NAC-type transcription factor OsNAC6 involved in abiotic and biotic stress-responsive gene expression in rice. *The Plant Journal*, 51, 617–630.
- Noguero, M., Atif, R. M., Ochatt, S., & Thompson, R. D. (2013). The role of the DNA-binding one zinc finger (DOF) transcription factor family in plants. *Plant Science*, 209, 32–45.
- Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Current Protocols in Keanean Journal of Science, Bioinformatics 2* (2013) 75, John Wiley & Sons, Inc. 42, 3.1.1–3.1.8.
- Perez-Rodriguez, P., Riano-Pachon, D. M., Correa, L. G. G., & Rensing, S. A. (2009). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research*.
- Pertsemelidis, A., & Fondon, J. W., III (2001). Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biology Reviews*, 2(10), 2002.1–2002.10.
- Pradhan, S., Kant, C., Verma, S., & Bhatia, S. (2017). Genome-wide analysis of CCCH zinc finger family identifies tissue specific and stress responsive candidate in chickpea (*Cicer arietinum* L.). *PLoS One*, 12(7), e0180469.
- Priya, P., & Jain, M. (2013). RiceSRTFDB: A database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant information to facilitate gene function analysis. *Database (Oxford)*, bat027.
- Pu, X., Yang, L., Liu, L., Dong, X., et al. (2020). Genome-wide analysis of the MYB transcription factor superfamily in *Physcomitrella patens*. *International Journal of Molecular Sciences*, 21(3), 975.
- Rashid, M., Guangyuan, H., Guangxiao, Y., Hussain, J., & Xu, Y. (2012). AP2/ERF transcription factor in rice: Genome-wide canvas and syntenic relationships between monocots and eudicots. *Evolutionary Bioinformatics*, 8, 321–355.
- Riechmann, J. L., Heard, J., Martin, G., Reuber, L., et al. (2000). Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. *Science (New York, N.Y.)*, 290(5499), 2105–2110.
- Romeuf, I., Tessier, D., Dardevet, M., et al. (2010). wDBTF: An integrated database resource for studying wheat transcription factor families. *BMC Genomics*, 11, 185.
- Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4), 725–738.
- Rushton, P. J., Bokowiec, M. T., Laudeman, T. W., Brannock, J. F., et al. (2008). TOBFAC: The database of tobacco transcription factors. *BMC Bioinformatics*, 9, 53.
- Salih, H., Gong, W., He, S., Sun, G., et al. (2016). Genome-wide characterization and expression analysis of MYB transcription factors in *Gossypium hirsutum*. *BMC Genetics*, 17, 129.
- Salih, H., Odongo, M. R., Gong, W., He, S., & Du, X. (2019). Genome-wide analysis of cotton C2H2-zinc finger transcription factor family and their expression analysis during fiber development. *BMC Plant Biology*, 19(1), 400.
- Sato, Y., Takehisa, H., Kamatsuki, K., Minami, H., et al. (2013). RiceXPro Version 3.0: Expanding the informatics resource for rice transcriptome. *Nucleic Acids Research*, 41, D1206–D1213.
- Sayers, E. W., Beck, J., Bolton, E., Bourexis, D., et al. (2021). *Nucleic Acids Research*, 49, D10–D17.
- Scharf, K. D., Berberich, T., Ebersberger, I., & Nover, L. (2012). The plant heat stress transcription factor (Hsf) family: Structure, function and evolution. *Biochimica et Biophysica Acta*, 1819, 104–119.
- Sebastian, A., & Contreras-Moreira, B. (2014). FootprintDB: A database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics (Oxford, England)*, 30(2), 258–265.
- Shameer, K., Ambika, S., Varghese, S. M., Karaba, N., et al. (2009). STIFDB-Arabidopsis stress responsive transcription factor database. *International Journal of Plant Genomics*, 583429.

- Shan, T., Rong, W., Xu, H., Du, L., et al. (2016). The wheat R2R3-MYB transcription factor TaRIM1 participates in resistance response against the pathogen *Rhizoctonia cerealis* infection through regulating defense genes. *Scientific Reports*, 6, 1–14.
- Shaw, L. M., McIntyre, C. L., Gresshoff, P. M., & Xue, G. P. (2009). Members of the Dof transcription factor family in *Triticum aestivum* are associated with light-mediated gene regulation. *Functional & Integrative Genomics*, 9(4), 485–498.
- Shen, S., Zhang, Q., Shi, Y., Sun, Z., et al. (2019). Genome-wide analysis of the NAC domain transcription factor gene family in *Theobroma cacao*. *Genes*, 11, 35.
- Sheng, X. G., Zhao, Z. Q., Wang, J. S., Yu, H. F., et al. (2019). Genome wide analysis of MADS-box gene family in *Brassica oleracea* reveals conservation and variation in flower development. *BMC Plant Biology*, 19, 106.
- Singh, A., Sharma, A. K., Singh, N. K., & Sharma, T. R. (2017). PpTFDB: A pigeonpea transcription factor database for exploring functional genomics in legumes. *PLoS One*, 12(6), e0179736.
- Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics (Oxford, England)*, 21(7), 951–960.
- Song, X., Nie, F., Chen, W., et al. (2020). Coriander Genomics Database: A genomic, transcriptomic, and metabolic database for coriander. *Horticulture Research*, 7(1), 55.
- Spivak, A. T., & Stormo, G. D. (2012). ScerTF: A comprehensive database of benchmarked position weight matrices for *Saccharomyces* species. *Nucleic Acids Research*, 40, D162–D168.
- Steffens, N. O., Galuschka, C., Schindler, M., Bülow, L., & Hehl, R. (2004). AthaMap: An online resource for *in-silico* transcription factor binding sites in the *Arabidopsis thaliana* genome. *Nucleic Acids Research*, 32, D368 = 372.
- Sugano, S., Kaminaka, H., Rybka, Z., Catala, R., et al. (2003). Stress-responsive zinc finger gene ZPT2–3 plays a role in drought tolerance in petunia. *The Plant Journal*, 36, 830–841.
- Sun, W., Ma, X., Chen, H., & Liu, M. (2019). MYB gene family in potato (*Solanum tuberosum* L.): Genome-wide identification of hormone-responsive reveals their potential functions in growth and development. *International Journal of Molecular Sciences*, 20(19), 4847.
- Tan, L., Ijaz, U., Salih, H., Cheng, Z., et al. (2020). Genome-wide identification and comparative analysis of MYB-transcription factor family in *Musa acuminata* and *Musa balbisiana*. *Plants*, 9(4), 413.
- Thamilarasan, S. K., Park, J. I., Jung, H. J., & Nou, I. S. (2014). Genome-wide analysis of the distribution of AP2/ERF transcription factors reveals duplication and CBFs genes elucidate their potential function in *Brassica oleracea*. *BMC Genomics*, 15, 422.
- Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., & Higgins, D. G. (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25, 4876–4882.
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22, 4673–4680.
- Tran, L.-S. P., Nakashima, K., Sakuma, Y., Simpson, S. D., et al. (2004). Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought responsive cis-element in the early responsive to dehydration stress 1 promoter. *The Plant Cell*, 16, 2481–2498.
- Tseng, K. C., Li, G. Z., Hung, Y. C., Chow, C. N., et al. (2020). EXPath 2.0: An updated database for integrating high-throughput gene expression data with biological pathways. *Plant and Cell Physiology*, 61, 1818–1827.
- Udvardi, M. K., Kakar, K., Wandrey, M., Montanari, O., et al. (2007). Legume transcription factors: Global regulators of plant development and response to the environment. *Plant Physiology*, 144(2), 538–549.
- Vannini, C., Locatelli, F., Bracale, M., Magnani, E., et al. (2004). Overexpression of the rice Osmyb4 gene increases chilling and freezing tolerance of *Arabidopsis thaliana* plants. *The Plant Journal*, 37, 115–127.
- Verma, S. K., Jasrotia, R. S., Iquebal, M. A., Jaiswal, S., et al. (2017). Deciphering genes associated with root wilt disease of coconut and development of its transcriptomic database (CnTDB). *Physiological and Molecular Plant Pathology*, 100, 255–263.
- Wang, D., Guo, Y., Wu, C., & Yang, G. (2008). Genome-wide analysis of CCCH zinc finger family in *Arabidopsis* and rice. *BMC Genomics*, 9, 44.
- Wang, R., Hong, G., & Han, B. (2004). Transcript abundance of rml1, encoding a putative GT1-like factor in rice, is up-regulated by *Magnaporthe grisea* and down-regulated by light. *Gene*, 324, 105–115.
- Wang, R., Zhao, P., Kong, N., Lu, R., et al. (2018). Genome-wide identification and characterization of the potato NHLH transcription factor family. *Genes*, 9, 54.
- Wang, Y., & Liu, A. (2020). Genomic characterization and expression analysis of basic helix-loop-helix (bHLH) family genes in traditional Chinese herb *Dendrobium officinale*. *Plants*, 9, 1044.
- Wang, Y., Zhang, J., Hu, Z., Guo, X., et al. (2019). Genome-wide analysis of the MADS-Box transcription factor family in *Solanum lycopersicum*. *International Journal of Molecular Sciences*, 20(12), 2961.
- Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics*, 2(4), 216–221.
- Wang, Z., Libault, M., Joshi, T., et al. (2010). SoyDB: A knowledge database of soybean transcription factors. *BMC Plant Biology*, 10, 14.
- Wang, Z., Yan, L., Wan, L., & Huai, D. (2019). Genome-wide systematic characterization of bZIP transcription factors and their expression profiles during seed development and in response to salt stress in peanut. *BMC Genomics*, 20, 51.
- Waqas, M., Azhar, M. T., Rana, I. A., Azeem, F., et al. (2019). Genome-wide identification and expression analyses of WRKY transcription factor family members from chickpea (*Cicer arietinum* L.) reveal their role in abiotic stress-responses. *Genes and Genomics*, 41, 467–481.
- Wei, K., & Chen, H. (2018). Comparative functional genomic analysis of bHLH gene family in rice, maize and wheat. *BMC Plant Biology*, 18, 309.
- Wei, K. F., Chen, J., Chen, Y. F., Wu, L. J., & Xie, D. X. (2012). Molecular phylogenetic and expression analysis of the complete WRKY transcription factor family in maize. *DNA research*, 19(2), 153–164.

- Wei, Q., Wang, W., Hu, T., Hu., et al. (2018). Genome-wide identification and characterization of Dof transcription factors in eggplant (*Solanum melongena* L.). *PeerJ*, 6, e4481. Available from <https://doi.org/10.7717/peerj.4481>.
- Wei, W., Huang, J., Hao, Y. J., Zou, H. F., et al. (2009). Soybean GmPHD-type transcription regulators improve stress tolerance in transgenic Arabidopsis plants. *PLoS One*, 4(9), e7209.
- Wingender, E., Dietze, P., Karas, H., & Knüppel, R. (1996). TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24, 238–241.
- Xia, L., Zou, D., Sang, J., Xu, X. J., et al. (2017). Rice Expression Database (RED): An integrated RNA-Seq-derived gene expression database for rice. *Journal of Genetics and Genomics*, 44(5), 235–241.
- Xie, Z., Nolan, T. M., Jiang, H., & Yin, Y. (2019). A2/ERF transcription factor regulatory networks in hormone and abiotic stress responses in Arabidopsis. *Frontiers of Plant Science*, 10, 228.
- Xie, Z. M., Zou, H. F., Lei, G., Wei, W., et al. (2009). Soybean Trihelix transcription factors GmGT-2A and GmGT-2B improve plant tolerance to abiotic stresses in transgenic Arabidopsis. *PLoS One*, 4(9), e6898.
- Xiong, J. (2006). *Protein motifs and domain prediction. Essential bioinformatics* (pp. 85–94). Cambridge: Cambridge University Press.
- Xu, Q. J., & Cui, C. R. (2007). Genetic transformation of OSISAPI gene to onion (*Allium cepa* L.) mediated by a microprojectile bombardment. *Journal of Plant Physiology and Molecular Biology*, 33, 188–196.
- Yadav, D., Malviya, N., Nasim, J., & Kumar, R. (2016). Bioinformatics intervention in elucidating structural and functional attributes of plant specific transcription factors: A review. *Research Journal of Biotechnology*, 11(7), 83–96.
- Yanagisawa, S. (2002). The Dof family of plant transcription factors. *Trends in Plant Science*, 7(12), 555–560.
- Yang, F., Li, W., Jiang, N., Yu, H., et al. (2017). Maize gene regulatory network for phenolic metabolism. *Molecular Plant*, 10(3), 498–515.
- Yang, J., Li, Y., Liu, Q., et al. (2020). Brief introduction of medical database and data mining technology in big data era. *Journal of Evidence-Based Medicine*, 13, 57–69.
- Yang, M. L., Chao, J. T., Wang, D. W., Hu, J. H., et al. (2016). Genome-wide identification and expression profiling of the C2H2-type zinc finger protein transcription factor family in tobacco. *Yi Chuan = Hereditas / Zhongguo yi Chuan xue hui Bian ji*, 38(4), 337–349.
- Yang, Q. Q., Feng, K., Xu, Z. S., Duan, A. Q., et al. (2019). Genome-wide identification of bZIP transcription factors and their responses to abiotic stress in celery. *Biotechnology and Biotechnological Equipment*, 33(1), 707–718.
- Yilmaz, A., Mejia-Guerra, M. K., Kurz, K., Liang, X., et al. (2011). AGRIS: Arabidopsis gene regulatory information server, an update. *Nucleic Acids Research*, 39(Database issue), D1118–D1122.
- Yogendra, K. N., Kumar, A., Sarkar, K., Li, Y., et al. (2015). Transcription factor StWRKY1 regulates phenylpropanoid metabolites conferring late blight resistance in potato. *Journal of Experimental Botany*, 66(22), 7377–7389.
- Yoo, S. D., Cho, Y. H., & Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature Protocols*, 2, 1565–1572.
- Yu, F., Liang, K., Fang, T., Zhao, H., et al. (2019). A group VII ethylene response factor gene, ZmEREB180, coordinates waterlogging tolerance in maize seedlings. *Plant Biotechnology Journal*, 17, 2286–2298.
- Yu, Y. H., Li, X. Z., Wu, Z. J., et al. (2016). VvZFP11, a Cys2His2-type zinc finger transcription factor, is involved in defense responses in *Vitis vinifera*. *Biologia Plantarum*, 60, 292–298.
- Zhang, C., Wang, D., Yang, C., Kong, N., et al. (2017). Genome-wide identification of the potato WRKY transcription factor family. *PLoS One*, 12(7), e0181573.
- Zhang, H., Pan, X., Liu, S., Lin, W., et al. (2021). Genome-wide analysis of AP2/ERF transcription factor in pineapple reveals functional divergence during flowering induction mediated by ethylene and floral organ development. *Genomics*, 113(2), 474–489.
- Zhang, X., Yao, C., Fu, S., Xuan, H., et al. (2018). Stress2TF: A manually curated database of TF regulation in plant response to stress. *Gene*, 638, 36–40.
- Zhao, P., Ye, M., Wang, R., Wang, D., & Chen, Q. (2020). Systematic identification and functional analysis of potato (*Solanum tuberosum* L.) bZIP transcription factors and overexpression of potato bZIP transcription factor StbZIP-65 enhances salt tolerance. *International Journal of Biological Macromolecules*, 161, 155–167.
- Zhao, T., Wu, T., Zhang, J., & Wang, Z. (2020). Genome-wide analysis of the genetic screening of C2C2-type zinc finger transcription factors and abiotic and biotic stress responses in tomato (*Solanum lycopersicum*) based on RNA-seq data. *Frontiers in Genetics*, 11, 540.
- Zhao, Y., Zhao, H., Wang, Y., Zhang, X., et al. (2020). Genome-wide identification and expression analysis of MIKC-type MADS-box gene family in *Punica granatum* L. *Agronomy*, 10, 1197.
- Zhang, Z., Li, J., Zhao, X. Q., Wang, J., et al. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, 4(4), 259–263. Available from [https://doi.org/10.1016/S1672-0229\(07\)60007-2](https://doi.org/10.1016/S1672-0229(07)60007-2).
- Zheng, X., Chen, B., Lu, G., & Han, B. (2009). Overexpression of a NAC transcription factor enhances rice drought and salt tolerance. *Biochemical and Biophysical Research Communications*, 379, 985–989.
- Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., et al. (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular Plant*, 9, 1667–1670.
- Zhou, Q. Y., Tian, A. G., Zou, H. F., Xie, Z. M., et al. (2008). Soybean WRKY-type transcription factor genes, GmWRKY13, GmWRKY21, and GmWRKY54, confer differential tolerance to abiotic stresses in transgenic Arabidopsis plants. *Plant Biotechnology Journal*, 6, 486–503.
- Zhu, L., Guo, J., Ma, Z., Wang, J., & Zhou, C. (2018). Arabidopsis transcription factor MYB102 increases plant susceptibility to aphids by substantial activation of ethylene biosynthesis. *Biomolecules*, 8, 39.
- Zhu, Q. H., Guo, A. Y., Gao, G., Zhong, Y. F., et al. (2007). DPTF: A database of poplar transcription factors. *Bioinformatics (Oxford, England)*, 23(10), 1307–1308.

Proteomics as a tool to understand the biology of agricultural crops

Riyazuddin Riyazuddin¹, Ashish Kumar Choudhary², Nisha Khatri³, Abhijit Sarkar⁴, Ganesh Kumar Agrawal⁵, Sun Tae Kim⁷, Ravi Gupta⁸ and Randeep Rakwal^{5,6,9}

¹Department of Plant Biology, Faculty of Science and Informatics, University of Szeged, Szeged, Hungary, ²Department of Botany, University of Delhi, Delhi, India, ³Department of Botany, Dyal Singh College, University of Delhi, Delhi, India, ⁴Department of Botany, University of Gour Banga, Malda, West Bengal, India, ⁵Research Laboratory for Biotechnology and Biochemistry (RLABB), Kathmandu, Nepal, ⁶GRADE (Global Research Arch for Developing Education) Academy Private Limited, Birgunj, Nepal, ⁷Department of Plant Bioscience, Pusan National University, Miryang, Republic of Korea, ⁸Department of Botany, School of Chemical and Life Sciences, Jamia Hamdard, New Delhi, India, ⁹Faculty of Health and Sport Sciences, University of Tsukuba, Tsukuba, Japan

7.1 Introduction

Because of climate change and associated effects of global warming, and an exponentially growing population (reaching around 7.8 billion), food security is now one of the major concerns of discussion globally. However, the past few decades have witnessed an increase in world food production because of the exploitation of different biological approaches. As per an estimate by the Food and Agriculture Organization, the future consumption of cereals will elevate by 70% by 2050 because of expanding populations and a shrinkage in the area of agricultural lands along with an increased severity of many (a) biotic stresses. Nowadays, feeding people with nutritious food and fulfilling their basic requirements of shelter and clothing at an affordable price has become a great challenge. Plants support all of these human requirements by providing food, fodder, fiber, and a framework for the shelter in the form of bamboo and wood derived from several other plants. Therefore agriculture is majorly based on fulfilling these human requirements and the topmost cultivated plants include rice, wheat, maize, soybean, tomato, mustard, grape, potato, sugarcane, and cotton (Fig. 7.1).

Rice, wheat, and maize account for half of the total calories consumed by the world's population (Maclean et al., 2002). Thus rice and other grain crops such as wheat, maize, and others require utmost consideration of scientists/breeders in both fundamental and applied research (Sarkar et al., 2014). However, the crops, including rice, wheat, and maize, are relatively more susceptible to the abiotic stresses which affect multiple aspects of plants' life cycle starting from the germination to seed development, and that is also reflected at the level of the proteome (Agrawal & Rakwal, 2006, 2011; Agrawal et al., 2006; Agrawal, Jwa, & Rakwal, 2009; Rakwal & Agrawal, 2003; Sarkar et al., 2014). Output results from the experimental data of model plants can be applied to agriculture crops to unravel the questions faced in the agriculture field (Agrawal et al., 2012; Jorin-Novo et al., 2009), including enhanced crop tolerance to environmental stresses and enhancing the quality in terms of nutrition values and yield of agricultural production to confirm food safety and security (Vanderschuren, Lentz, Zainuddin, & Gruissem, 2013).

Proteomics refers to the comprehension of the global protein expression in an organism (Aebersold & Mann, 2003; Gupta, Wang, et al., 2015). Identifying and analyzing the proteins and their expression in different organisms under different physiological conditions helps plant biologists to better understand the adaptive response of those organisms under that particular stress. Proteomics is a rapidly growing and advancing field of science as a better understanding of the proteins enables one to better understand the complex metabolic processes, protein interaction, and the regulatory pathways, which later can be altered according to our needs. In the case of plant science, understanding the proteins involved and their interaction in the adaptive response to various abiotic and biotic stresses is useful in creating

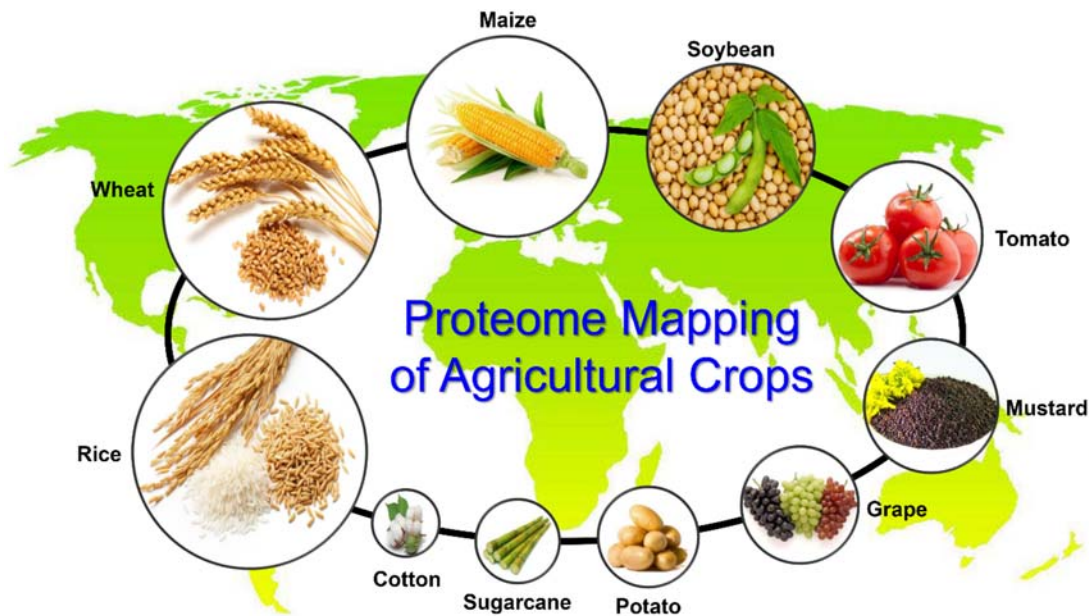


FIGURE 7.1 The top-10 valued agricultural crops where proteomics has been utilized. The size of each crop represents the total number of proteomics studies published on that particular crop.

transgenic plants better adaptive to the external environmental stresses. In the case of agricultural crops, the majority of the proteomics studies have been carried out on the rice (Kim, Kim, Agrawal, Kikuchi, & Rakwal, 2014; Meng et al., 2019; Vanderschuren et al., 2013), wheat (Komatsu, Kamal, & Hossain, 2014), maize (Pechanova, Takáč, Šamaj, & Pechan, 2013), soybean (Min, Gupta, Agrawal, Rakwal, & Kim, 2019), and tomato (Ghatak et al., 2017) because of the obvious reasons of their consumption (Fig. 7.2A) and the number of proteomics studies on these crops have progressively increased in the last decade (Fig. 7.2B) because of the advancements in proteomics technologies. This knowledge from these proteomics studies has been applied to some extent to improve the agricultural output and reduce the loss caused due to abiotic or biotic stresses. The tools applied to study proteins are either gel-based techniques or gel-free analysis. Gel-based techniques include two-dimensional gel electrophoresis (2DGE) or difference-in-gel electrophoresis (DIGE) for protein separation and quantification and, on the other hand, gel-free techniques omit the separation of proteins on the gels and the isolated proteins are directly subjected to the protein digestion followed by liquid chromatography (LC) and then mass spectrometry (MS) identification (Champagne & Boutry, 2013). Gel-based techniques are usually used to get a visual pattern of the protein changes in response to an external or internal stimulus on the polyacrylamide gels. In a gel-free technique, either labeling followed by identification is followed or procedures such as multidimensional capillary LC coupled to nano-ESI tandem MS to separate, and then identification is followed (Riter et al., 2011). However, both techniques have their limitations. For example, 2DGE which is used for separation can only separate about 30%–40% of the total proteome also, strongly alkaline proteins with $\text{pH} > 9.5$ are also difficult to focus (Chevalier, 2010). The limitations of gel-based techniques led to the advancement of gel-free techniques. Gel-free procedures should be chosen according to the requirement and aims of the experiment and the sample being used. The latter technique is better as they are more reproducible and accurate than the former. Advancements in proteomics approaches and recent progress in the plant proteomics research have been elegantly discussed previously (Agrawal & Rakwal, 2008).

7.2 Gel-based proteomics

Gel-based proteomics remained the method of choice for proteome analysis of plant samples in the last two decades (Liu et al., 2019; Righetti, 2014). It primarily involves the resolution of proteins on gels prior to their trypsin digestion and identification by MS (Fig. 7.3). For protein separation, multiple options, including sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) or one-dimensional gel electrophoresis, 2DGE, and DIGE, are available of which 2DGE has been primarily utilized for the plants (Fig. 7.3). In general, gel-based proteomics involves the resolution of isolated proteins from control and treated samples on polyacrylamide gels that separates the proteins based

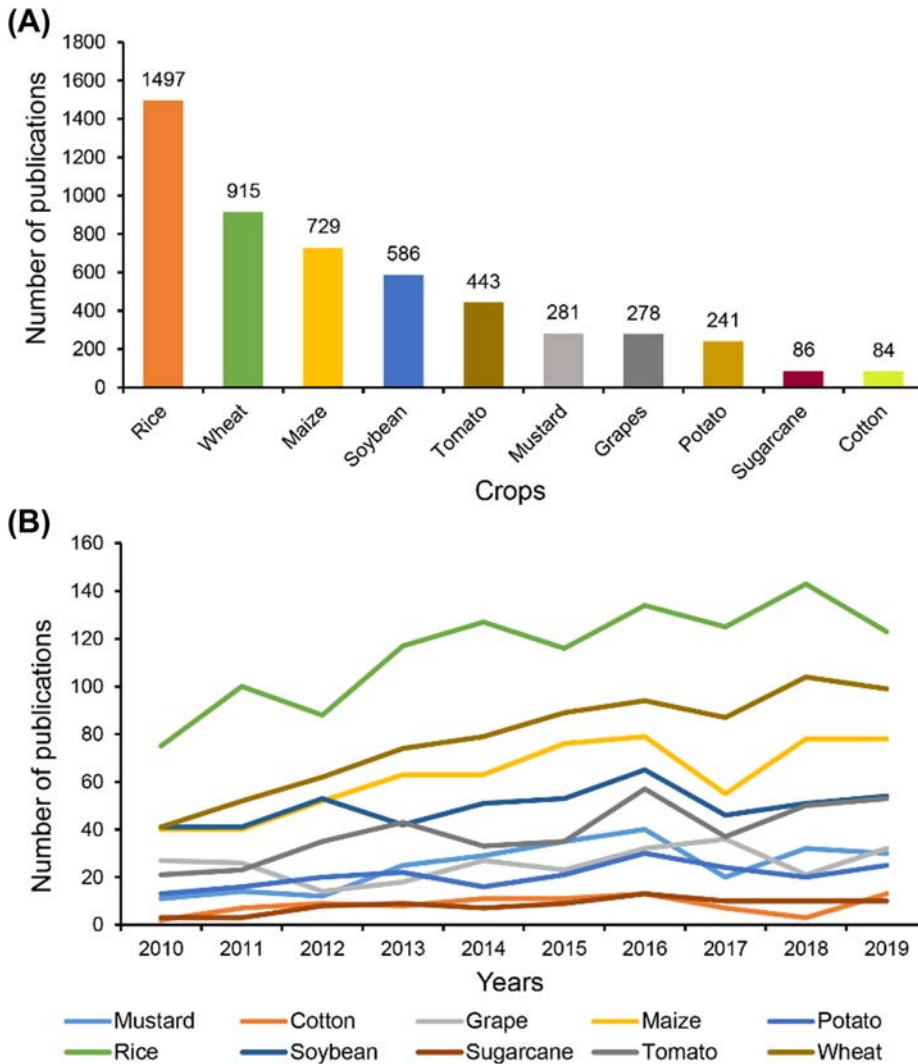


FIGURE 7.2 (A) Number of proteomics studies carried out to date in different crop plants. (B) An overview of the proteomics studies carried out in different crops in the last 10 years. Data were retrieved from the PubMed on July 28, 2020.

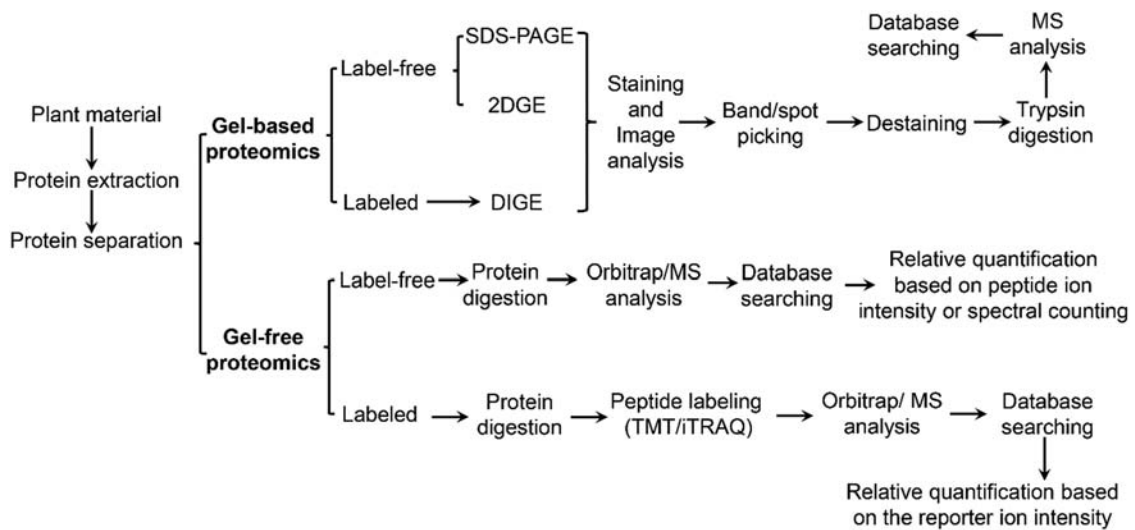


FIGURE 7.3 An overview of the proteomics technologies utilized so far for the comparative proteome analysis of the crop plants.

on their molecular weight and/or charge. The gels are then stained with the protein staining dyes to visualize the protein bands as in the case of SDS-PAGE or protein spots in the case of 2DGE and 2D-DIGE. The obtained gel pictures are digitized and the differential expression of proteins is determined with the help of sophisticated software. Identified differential proteins are excised from the gels, destained, and digested by proteases, most commonly with the trypsin to generate a set of peptides that are then identified by MS (Fig. 7.3). Just in the case of rice and wheat, more than 950 studies have been published on the 2DGE analysis to decipher the molecular mechanisms of the stress response, growth, and development, highlighting the central role of gel-based proteomics in the crop proteomics (Agrawal et al., 2013; Komatsu et al., 2014).

7.2.1 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis

Development of SDS-PAGE by Laemmli (1971) was a major breakthrough in the life science research which involves the separation of proteins based on their molecular weight on a polyacrylamide gel. Each protein in this method is labeled with the SDS which imparts a negative charge to all the proteins rendering their usual charge insignificant and therefore protein separation takes place solely based on their sizes. This method involves first the reduction of all the disulfide bonds in the isolated proteins either by DTT or β -mercaptoethanol followed by boiling to break down the proteins into corresponding polypeptides that appear as bands on the SDS-PAGE after the staining. However, the resolution of the SDS-PAGE is limited and each band is often a mixture of several polypeptides of the same size. In the case of proteome analysis, the only way to use SDS-PAGE is the resolution of polypeptides on the gel followed by cutting each sample lane in 5–6 segments, all of which are then subjected to in-gel trypsin digestion followed by protein identification by MS. This method of proteome analysis has been used majorly in rice to identify the different fungal and bacterial responsive proteins in the apoplast and symplast (Wang et al., 2017). However, since this method alone cannot be used for the quantification of proteins, it is often combined with the label-free quantitative proteomics to identify the stress-responsive proteins in a high-throughput manner (Gupta et al., 2016a, 2016b). The benefit of using this method is the fractionation of proteins prior to their digestion to reduce the complexity of peptides that hinder protein identification by MS.

7.2.2 Two-dimensional gel electrophoresis

2DGE was first introduced by three scientists, including Klose, O'Farrell, and Scheele independently in 1975 (Klose, 1975; O'Farrell, 1975; Scheele, 1975). 2DGE involves the resolution of proteins in two dimensions where the first separation is based on the isoelectric point or charge of the proteins while the second-dimensional separation is based on the molecular weight. Immobilized pH gradient strips have been developed for the separation of proteins based on their charge and are available in a variety of configurations and sizes as per the requirement (Bjellqvist et al., 1982). These strips are available as pH 3–7 nonlinear and linear gradient and 4–7 linear gradient starting from 7 cm to as large as 24 cm. In contrast, the second-dimensional separation can be carried out as a routine SDS-PAGE. This two-dimensional separation of proteins separates different proteins as individual spots that can be visualized using a variety of staining techniques, including silver staining, Coomassie brilliant blue (CBB), and colloidal CBB, among others. Besides, fluorescent stains such as Sypro Ruby have also been developed with improved sensitivity than CBB and colloidal CBB. However, the silver staining method is the most sensitive method developed so far for the detection of proteins on the gels and is able to detect proteins even in nanograms (Wray, Boulikas, Wray, & Hancock, 1981). This 2DGE-based framework provides a snapshot of the total proteome of tissue at a given time and permits an immediate examination of the treated sample with control (Gupta, Min, et al., 2015). This method provides accurate and reliable differences between stress-treated samples with the control and includes quantification of proteins on the foundation of the spot intensities. A single 2DGE gel can resolve up to 10,000 spots representing 1000 proteins, highlighting its excellent resolution power (Abdallah, Dumas-Gaudot, Renaut, & Sergeant, 2012). However, 2DGE has limitations in protein solubilizing, reproducibility, detection of low-abundant proteins, and resolution of highly acidic or basic proteins (Anguraj Vadivel, 2015), yet it has remained the method of choice for the analysis of plant proteomes, especially rice and wheat, to identify the differential proteins in response to various biotic and abiotic stress conditions.

7.2.3 Two-dimensional-difference-in-gel electrophoresis

2D-DIGE, a variant of 2DGE, was first introduced by Unlu et al. (1997), 22 years later than the introduction of classical 2DGE. 2D-DIGE involves the separation of control and treated samples on the same 2DGE gel, thereby reducing the

gel-to-gel variations, which is a major drawback of the classical 2DGE system. DIGE is based on the labeling of control and treated samples with fluorescent dyes, named CyDyes. The NHS-ester reactive group of these dyes covalently attaches to the ϵ -amino group of lysine residues of the proteins via an amide linkage. At present, there are three CyDyes available named Cy2, Cy3, and Cy5 in which the control sample is usually labeled with Cy3, treated with Cy5, and pooling of both the samples as an internal standard with Cy2. The internal standard helps in spot normalization and provides a snapshot of all the resolved proteins from control and treated samples. Since these are the fluorescent dyes, the spots in the gels can only be visualized using a fluorescence imager such as Typhoon variable mode imager by GE Healthcare. Despite the significant advancements in the gel analysis and spots detection, DIGE has been poorly utilized for the analysis of proteomes of rice and wheat. In case of rice, DIGE has been used for the comparative proteome analysis of different rice cultivars (Teshima, Nakamura, Satoh, & Nakamura, 2010), low-level gamma radiation (Hayashi et al., 2015; Rakwal et al., 2018), heat-responsive proteins (Zhou et al., 2019), analysis of secreted proteins in response to salt stress (Song et al., 2011), and *Xanthomonas oryzae* pv. *oryza* infection (Chen, Deng, Yu, Yan, & Chen, 2016).

7.3 Gel-free proteomics

7.3.1 Multidimensional Protein Identification Technology

Multidimensional Protein Identification Technology (MudPIT) is a gel-free relatively advanced technique that was developed in the Yates Laboratory. MudPIT is an high-performance liquid chromatography (HPLC)-based peptide separation approach coupled with MS which requires digested protein samples into peptides before the separation steps (Washburn, Wolters, & Yates, 2001). This technique works based on the combination of strong cation exchange and reverse-phase chromatography to accomplish two-dimensional separation prior to MS analysis (García, Senis, Tomlinson, & Watson, 2007). To overcome the issue with 2-DE, MudPIT was introduced which is a robust and broadly recognized method for protein recognition from a wide variety of samples. This gel-free technique is capable of separating individual components of protein and peptides from complex mixtures, resulting in a high number of protein identifications, including low abundant ones. MudPIT is an excellent tool for both qualitative and quantitative proteomic analyses and is mainly appropriate for the identification of hydrophobic proteins (Rossignol et al., 2006). However, its use in plants, especially for the crops, is still limited. At first, Koller et al. (2002) reported the use of both MudPIT and 2DGE to characterize the proteome of different rice tissues, including leaves, roots, and seed tissues of rice. Further, the MudPIT approach was also applied to identify potential targets to prevent the population expansion of phytophagous mite in rice plants (Blasi et al., 2017). However, they also identified the expression pattern of proteins in the infested leaves of rice, and results indicate that the acceptor side of photosystem II is probably the major susceptible target in the photosynthetic apparatus (Buffon et al., 2016). In addition, Maor et al. (2007) applied the MudPIT approach to investigate the protein ubiquitination in various plant species (Maor et al., 2007).

7.3.2 Sequential window acquisition of all theoretical mass spectra

The term “sequential window acquisition of all theoretical mass spectra (AP-SWATH)” was first used by Gillet et al. (2012). AP-SWATH is relatively a recent technique based on data-independent acquisition, introduced to extend the proteome coverage of shotgun proteomics (Tsou et al., 2015). In principle, the SWATH method can perform label-free quantification in a multiple reaction monitoring such as fashion, which has higher quantification accuracy and precision. AP-SWATH exhibits a high rate of accuracy, sensitivity, and reproducibility for evaluating the whole proteome. In contrast to other techniques, the main requirement for this method is the creation of a comprehensive spectral library (Yin, He, Gupta, & Yang, 2015). Since it is a relatively recently introduced technique, its application to the crop species is yet limited. SWATH-MS approach was used in different rice varieties to check the quality and proteins associated with nutritional values (Sew et al., 2020). Moreover, the SWATH-MS approach was also employed for analysing proteins quantitatively which were present in both inferior and superior spikelets of rice plants in the course of the grain filling period (Zhu et al., 2016). During the germination of rice, Zhang, Wang, et al. (2016) studied the dynamic protein carbonylation and found the involvement of carbonylated proteins in reactive oxygen species (ROS) homeostasis, stress hormones, and seed reserves via AP-SWATH label-free technique. To examine the protein profiles of a variety of japonica rice plants during grain filling stages under normal as well as under stress (moderate soil drying) conditions, the SWATH-MS method was used. Results obtained suggested that the remobilization of rice straw carbon reserve was a mainly due to the altered gene expression during grain filling period in response of moderate soil drying conditions (Wang et al., 2020).

7.3.3 Label-free quantification

Currently, there are two main approaches to analyze the relative abundance of proteins in two or more biological samples. The first approach is label quantification in which the target proteins are labeled with a stable isotope whereas second approach include label free quantification that works on the principle of peptide intensity profiling. To calculate the relative abundance of the target modified peptide, the area under its peak is integrated and then is divided by the total area of the target peptide in unmodified and all modified forms. In label-free quantification, there are two commonly used quantitative schemes: (1) mass spectral peak intensities and (2) spectral counting and it has been observed that the amount of protein correlates well with peak intensities or spectral counts of peptides unique to a specific protein (Fig. 7.3). Li, Chen, He, and Yang (2020) applied a label-free quantification technique to investigate and characterize the importance of Ca^{2+} ions in rice seed germination. Label-free proteomic analysis was applied to organic and conventional rice (Meng et al., 2018; Wang et al., 2017; Xiao, Li, & Ma, 2019) and soybean samples (Gupta, Min, Kramer, et al., 2018). The label-free qualitative proteomics showed that photosynthesis-related pathways and ribosomal pathways were significantly inhibited in *OsCpn60β1* mutants of rice plants (Wu et al., 2020). To compare protein expression patterns in anthers during development, the relative quantity of proteins was estimated based on spectral counts, which is a label-free method (Bridges et al., 2007; Zhu, Smith, & Huang, 2010). The 2019 study investigated temporal changes in the abundance of the apoplastic fluid proteome of resistant and susceptible wheat leaves infected with *Puccinia triticina* race-1, using a label-free LC-MS-based approach (Rampitsch & Huang, 2019). The main advantages of label-free quantification are that it is cheap, achieves high-proteome coverage, and does not require laborious labeling workflows.

7.3.4 Isobaric tags for relative and absolute quantitation

Isobaric tags for relative and absolute quantitation (iTRAQ) is a technique of peptides labeling that uses isobaric reagents yielding amine-derivatized peptides. The reagents were developed by Darryl Pappin and colleagues at Applied Biosystem (Ross et al., 2004). Utilizing peptide fragments and low mass reporter ions, this approach allows quick protein identification and relative quantification at the MS/MS level (Evans et al., 2012). Chances of peak overlapping while analyzing the results are reduced due to the unique design of isobaric mass tags. During MS/MS analysis of iTRAQ-tagged peptides, the mass balancing carbonyl moiety is released as a neutral fragment and remaining isotope-encoded reporter ions are used for the relative quantification of protein (Fig. 7.3). Because of the availability of up to eight different iTRAQ reagents, a comparative analysis of multiple samples can be carried out in a single MS run (Wiese, Reidegeld, Meyer, & Warscheid, 2007). In plant science, this technique is used extensively and frequently to study adaptive stress responses in plants. For example, iTRAQ proteome was used to study the mechanism of ethylene-dependent salt response in bread wheat (Ma, Shi, Su, & Liu, 2020). Moreover, the technique was useful in analyzing new metabolic pathways of wheat seedling which were grown under hydrogen peroxide stress (Ge et al., 2013). In the case of rice, the technique was used to quantitatively compare the rice leaves of noninfected and infected plants (Wang, Ren, Lu, & Wang, 2015). Also, the technique was useful in providing insight into the molecular mechanism of cold-tolerance response in japonica rice (Jia et al., 2020).

7.3.5 Tandem mass tag

The use of tandem mass tag (TMT) serves to be a novel strategy for the accurate quantification of peptides and associated proteins. These new tags and their analysis protocols allow peptides from different samples to be identified based on their relative abundance. This method allows the simultaneous identification and estimation of relative abundances of peptides or proteins, facilitated by collision-induced dissociation-based analysis (Schlosser & Lehmann, 2000). This MS/MS-based detection protocol is known to have wider applications in peptide isolation methods when compared to the MS mode measurement, and the same is justified by the high signal-to-noise ratio achieved in MS/MS-based approach, allowing the exclusion of untagged materials, which helps in greatly improving the data quality. The TMT-based technique is analogous to other peptide isotope labeling techniques but offers additional advantages such as providing more precise reciprocal internal standards, which leads to more accurate quantification (Gupta et al., 2020; Gupta, Min, Kim, & Kim, 2019; Min et al., 2020). Also, the reagents used to apply to any peptide isolation protocols for in vitro labeling techniques. Robustness of the analysis and sensitivity of the TMT-based protein analysis help in obtaining efficient and accurate results when compared to the orthodox isotope labeling methods (Thompson et al., 2003). Liu et al. (2018) used the technique to analyze the response of tea plant to fluoride stress. Moreover, this

technique was also applied to study chlorophyll synthesis and chloroplast structure in *Brassica napus* (Yang et al., 2020). Moreover, Wu, Mirzaei, Pascovici, Haynes, and Atwell (2019) studied proteins from drought-resistant and drought-sensitive rice plants using TMT to better understand the effects of soil drying on gene expression.

7.3.6 Stable Isotope Labeling by Amino acids in Cell Culture

It is another technique used for comparative analysis of the cellular proteome. SILAC stands for Stable Isotope Labeling by Amino Acids in Cell Culture and was developed by Matthias Mann, of the University of Southern Denmark, and colleagues in the year 2002. It is a method of metabolic labeling using stable isotope. The isotope-labeled amino acids are mixed in the cell culture which is then incorporated in the proteins of the live cells (Ong et al., 2002). Since only selected amino acids are labeled, the quantification and comparison of the incorporated amino acids into the proteins is easier and accurate as well. For example, natural isotope and heavy isotope medium are allowed to grow for a few generations depending upon the protein formation and degradation process. After a few cycles, all the proteins in light, medium with natural isotope amino acids will have the natural isotope only, whereas in case of heavy, medium with stable isotope-labeled amino acids will have the proteins with the heavy isotope. The complex protein samples are then digested into peptides and analyzed. The signal intensities from light and heavy samples allow for a quantitative comparison of their relative abundances in the mixture (Chen, Wei, Ji, Guo, & Yang, 2015). For labeling of amino acids, the ones which are essential amino acids are preferred making sure that the only source of these amino acids is from the provided media. Leucine (Foster, De Hoog, & Mann, 2003; Ong et al., 2002), lysine (Everley, Krijgsveld, Zetter, & Gygi, 2004), and methionine (Ong, Mittler, & Mann, 2004) have been previously used providing promising results in SILAC labeling. The technique is quantitatively accurate, reproducible, and can even be used to analyze changes of posttranslational modifications (PTMs) and protein turnover (Chen et al., 2015). The applications of the technique are diverse and therefore it has also been used to identify protein effectors in the wheat-*Fusarium graminearum* pathosystem (Lecomte et al., 2014).

7.4 High-throughput posttranslational modification proteomics

Proteins are synthesized through translation from mRNAs in the form of nascent polypeptide chains on ribosomes. After translation, many proteins undergo chemical modifications or PTMs to form mature proteoforms that accumulate in the cells (Smith et al., 2013). Generally, these modifications take place in the Golgi apparatus and endoplasmic reticulum. PTMs are noticed throughout the life cycle of proteins. In all eukaryotic cells, more than 50% of total proteins at some point in their life cycle undergo PTMs (Cruz, Nguyen, Nguyen, & Wallace, 2019).

There are many forms of PTMs such as the addition of chemical groups (methylation, phosphorylation, and acetylation), the addition of complex molecules (AMPylation, glycosylation, and ADP-ribosylation), the addition of polypeptides (SUMOylation and ubiquitination), direct modification of amino acids (elimination and deamidation), and cleavage of protein by proteolytic mechanisms (Spoel, 2018). For example, 461 types of modified amino acids of eukaryotic proteins are deposited in the UniProt database (Bateman, et al., 2017) and newly identified PTMs are continuously being added to the list. These modifications of proteins affect their stability, activity, interactions, and the localization (Bateman et al., 2017). Proteins can exhibit different types of PTMs, including phosphorylation, acetylation, glycosylation, acylation, ADP-ribosylation, amidation, proteolytic processing, sulfation, disulfide bond formation, methylation, ubiquitylation, nitrosylation, sumoylation, γ -carboxylation, and β -hydroxylation (Villafañez, Gottifredi, & Soria, 2019; Walsh & Jefferis, 2006). These PTMs can be broadly classified into four groups based on their activities, namely, the addition of a peptide or protein, addition of a functional group, structural modification of a protein, and a change in the chemical nature of its amino acids (Ytterberg & Jensen, 2010). PTMs can either be of just one type or can be of various combinational types of modification that are highly specific and regulated to cellular requirements. These modifications are mainly specific to some cellular signals and are spatial (dependent on its location) and temporal (time-dependent) in nature. These modifications may also be influenced by the developmental stage and various types of abiotic and biotic factors. Some PTMs are irreversible for the lifetime of a protein such as glycosylation or cleavage of a signal peptide, while some modifications are reversible and quick in nature, for example, phosphorylation (Webster & Thomas, 2012). PTMs are very diverse and can regulate the function of proteins and their interacting partners through controlling the protein–protein interactions (Duan & Walther, 2015). It activates various biosynthetic pathways, namely, biotinylation of carboxylases, phosphopantetheinylation of fatty acid synthase, lipoylation of α -keto acid dehydrogenase, polyketide synthases, and nonribosomal peptide synthetases (Perham, 2000). These enzymatic activities play a major role in the regulation of various biological pathways for the proper functioning of cells.

Plants require very quick cellular sensing mechanisms because they are immobile and persistently exposed to various environmental changes (Dai Vu, Gevaert, & De Smet, 2018). Therefore various PTMs regulate metabolic pathways and functional traits during adverse conditions, which have been confirmed through a comprehensive analysis of plant quantitative proteomics data (Friso & Van Wijk, 2015). PTMs also enhance tolerance against several abiotic stresses, including salinity (Forment, Naranjo, Roldán, Serrano, & Vicente, 2002), freezing (Kim, Kim, & Kang, 2005), oxidative (Sunkar, Kapoor, & Zhu, 2006), cold (Kim et al., 2007), drought (Ko, Yang, & Han, 2006), heat, and osmotic (Kant, Kant, Gordon, Shaked, & Barak, 2007) stresses. Interestingly, most therapeutic proteins may also undergo several PTMs for their stability, pharmacokinetics, solubility, and bioactivity (Gomord & Faye, 2004). Various PTM-related databases have been generated from the information of Ms experiments of many model organisms, for example, PhosphoELM (Via et al., 2010), PHOSIDA (Gnad, Gunawardena, & Mann, 2010), PhosphoGRID (Sadowski et al., 2013), dbPTM (Kao et al., 2015), and iPTMNet (Huang et al., 2017). All of the developed proteomics techniques in principle can be utilized for the detection of PTMs in crops, including both gel-based and gel-free approaches (Hashiguchi & Komatsu, 2016).

7.4.1 Phosphorylation

Phosphorylation of proteins is a reversible attachment of phosphate group(s) to the serine, threonine and tyrosine of the proteins by the activity of kinases. In other words, protein phosphorylation is a reversible PTM of protein that regulates the important events of plants' life such as cellular signaling processes under normal as well as stress conditions (Gupta, Min, Meng, Agrawal, et al., 2018; Gupta, Min, Meng, Jun, et al., 2018). In phosphorylation, phosphoryl group from ATP (also from ADP) is transferred to the hydroxyl group of serine, threonine, or tyrosine residues of their target protein and these reactions are catalyzed by protein kinases (Champion, Kreis, Mockaitis, Picaud, & Henry, 2004). In the reverse process, the phosphorylated residues of modified proteins are removed by phosphatase activity (Friso & van Wijk, 2015). Hydroxylated amino acids such as serine (75%–80%), threonine (15%–20%), and tyrosine (1%–5%) majorly participate in phosphorylation of proteins (Champion et al., 2004). Apart from these hydroxylated amino acids, histidine and aspartic acid can also phosphorylate (Ciesla, Frączyk, & Rode, 2011). The protein kinases and phosphatases are the regulators of phosphorylation and these gene families are abundant in plants. For instance, Arabidopsis genome encodes for 162 phosphatases and 1052 protein kinases (Wang et al., 2014), undeniably showing the significance of phosphorylation. Mammalian genomes encode half of the protein kinases compared to plant genomes (Zulawski, Braginets, & Schulze, 2013). The role of protein kinase and phosphatase gene families in the regulation of phosphorylation have been extensively studied in crop plants such as wheat and maize (Singh, Giri, Kapoor, Tyagi, & Pandey, 2014; Wang et al., 2016; Wei & Pan, 2014).

Protein phosphorylation is essential for the proper regulation of photosynthesis in plants (Kwon, Choi, Choi, Ahn, & Park, 2006). Several other biological processes such as development differentiation, intracellular regulation, and cell maintenance are also dependent on phosphorylation and kinase activities (Vandamme, Castermans, & Thevelein, 2012). Phosphorylation regulates most of the physiological and metabolic pathways in plants, for example, RNA metabolism (van Bentem et al., 2006), defense (Jones, Bennett, Mansfield, & Grant, 2006; Nühse, Bottrill, Jones, & Peck, 2007), root growth (Zhang et al., 2013; Zhang, He, et al., 2016), and carbon metabolism (Wu, Sklodowski, Encke, & Schulze, 2014). The major function of protein phosphorylation is in the metabolic and signal transduction pathways by the alteration of protein activities such as protein interactions, subcellular localization, or protein (Mithoe & Menke, 2011; Schönberg & Baginsky, 2012; Silva-Sanchez, Li, & Chen, 2015; van Wijk, Friso, Walther, & Schulze, 2014). Phosphorylation also plays an important role during abiotic and biotic stresses. Many phosphorylation experiments have been performed under salt and drought stresses on crop plants such as wheat, sugar beet, maize, and several beans (Kumar et al., 2014; Lv et al., 2016; Yu et al., 2016; Zhang et al., 2014; Zörb, Schmitt, & Mühling, 2010). In the case of biotic stress such as host–pathogen interaction, it influences plant survival (Gupta et al., 2018; Xing, Ouellet, & Miki, 2002). Due to the enormous function of this PTM, several databases such as PhosPhat (Zulawski et al., 2013) and P3DB (Yao et al., 2014) have been specifically generated to store and retrieve phosphorylation information. These databases are very important for the understanding of phosphorylation mechanisms in the model and nonmodel plants.

7.4.2 Glycosylation

Glycosylation of protein is the most common and widespread PTM in plants and all other eukaryotes. It is an essential cotranslational modification and PTM that occurs in membrane and secreted proteins. In this chemical modification, the sugar molecule adds to the specific amino acids of the protein through enzymatic activity. This is found in natural

proteins as well as in biopharmaceutical proteins, for example, around 50% proteins are glycosylated in humans (Wong, 2005). As per the Swiss-Prot protein database, more than 50% of all eukaryotic proteins and one-third biopharmaceuticals protein are glycoproteins (Apweiler, Hermjakob, & Sharon, 1999; Walsh & Jefferis, 2006). Glycosylation is of two types and is classified based on the bond between the amino acid and carbohydrate (glycan) residues, O- and N-glycosylations. In N-glycosylation, N-glycans attached to the amide group of asparagine and O-glycans, O-glycosylation, attached to the hydroxyl group of serine, threonine, hydroxyproline, or hydroxylysine residues in the protein chain (Gomord et al., 2010; Williams, 2006). N-glycosylation starts in the endoplasmic reticulum and processing occurs in the Golgi apparatus while O-glycosylation occurs in the Golgi apparatus and endoplasmic reticulum (Gomord et al., 2010). N-Glycosylation plays a pivotal role in the regulation of several biological processes such as protein–protein interactions, protein folding, protein stability, and glycan-dependent activities in the endoplasmic reticulum (Hebert, Lamriben, Powers, & Kelly, 2014; Moremen, Tiemeyer, & Nairn, 2012). In *Arabidopsis thaliana*, more than 1000 N-glycosylated proteins have been noticed (Song et al., 2013; Zielinska, Gnad, Schropp, Wisniewski, & Mann, 2012) and these proteins contain one or several N-glycan.

Based on findings from *A. thaliana* (Strasser, Altmann, Mach, Glössl, & Steinkellner, 2004; von Schaewen, Sturm, O'Neill, & Chrispeels, 1993), it has been hypothesized that complex N-glycans are not essential for the development and reproduction of plants when grown under normal environmental conditions. Some investigations suggested that the N-glycans are not essential for the reproduction and development in *A. thaliana* under optimum environmental conditions (Strasser et al., 2004; von Schaewen et al., 1993). However, N-glycan modifications are conserved across land plants—from flowering plants to distantly related mosses, for example, *Physcomitrella patens* (Fitchette et al., 1999; Viëtor et al., 2003; Wilson et al., 2001). N-glycosylation plays an important role in many abiotic stresses.

The β 1,2-xylosyltransferase enzyme catalyzes the transfer of xylose subunit to the N-glycans but the loss of function of this enzyme causes impaired root aerenchyma formation. The nonfunctional β 1,2-xylosyltransferase enzyme in rice mutant is susceptible to osmotic stresses and low heat (Takano et al., 2015). Similarly, the α 1,3-fucosyltransferase enzyme catalyzes the transfer of fucose to the N-glycan cores. The α 1,3-fucosyltransferase enzyme in mutant rice showed a reduced gravitropic response because this enzyme is essential for the transfer of fucose to the N-glycan (Harmoko et al., 2016). The N-glycosylation affects several aspects which include plant metabolism, development, growth, stresses (biotic and abiotic responses), and enzyme functions (Friso & van Wijk, 2015). As a comparison to N-glycosylation, O-glycosylation has not been properly understood. This protein modification is mostly found in the cell wall and nearly 10% of cell wall proteins are O-glycosylated (Friso & van Wijk, 2015). In plants, two O-GlcNAc transferase enzymes (namely, SPINDLY and SECRET AGENT) regulate O-glycosylation, the double mutants of these enzymes (not functional) showed embryo lethality (Friso & van Wijk, 2015).

7.4.3 Acetylation

Acetylation is another important PTM of proteins. In acetylation modification, the acetyl group of acetyl-CoA is transferred to the amino acid (lysine) of protein by acetyltransferase enzymatic activity. This type of modification has been noticed in proteins generally found in all kinds of organisms, across all kingdoms, and therefore has been speculated to be universal (Choudhary, Weinert, Nishida, Verdin, & Mann, 2014; Jeffers & Sullivan, 2012). Protein acetylation can be both reversible and nonreversible and is mainly involved in the regulation of gene expression. For instance, ϵ -amino group modification of lysine (K) is a reversible process catalyzed by K acetyltransferases and K deacetylases, while N α -terminal modification is nonreversible process catalyzed by Nt-acetyltransferases (Nallamilli et al., 2014). In living organisms, more than 80% of proteins undergo N-terminal acetylation (Bienvenut et al., 2012) but the functional relevance of this PTM is not been properly understood so far. Previous investigations suggest that N-terminal acetylation is a fundamental regulator of several aspects of plant development, growth, behavior, protein stability, and stress response (Linster et al., 2015; Xu et al., 2015). Some plant growth regulators are up- and downregulators of this PTM, for example, gibberellin stress hormone promotes the modification of histones through the acetylation and hence suppresses the abscisic acid under note that both of these abiotic stresses (Hou et al., 2015). Linster et al. (2015) observed that downregulation of N-terminal acetylation promotes abscisic acid level and thus provides the adaptation against drought stress. The N-terminal acetylated protein in the *Arabidopsis* mutant lines showed growth defects and pleiotropic development (Ferrández-Ayela et al., 2013) with low photosynthetic capacity (Pesaresi et al., 2003).

Acetylation and deacetylation processes have been reported in the genome of several plant species and both processes are regulated by acetyltransferase and deacetylase enzymatic activities, respectively. *Arabidopsis* genome contains 12 genes for histone acetyltransferases and 18 genes for histone deacetylases (Pandey et al., 2002). As a comparison, 7 histone acetyltransferase and 19 histone deacetylase genes are present in the rice genome (Hu et al., 2009). Both histone and

nonhistone proteins undergo acetylation; however, this modification is controlled by many factors, such as ROS, physiological stresses, and infectious diseases. Histone acetylation regulates several significant cellular processes and metabolic activities such as development, flowering time, cell cycle, and signal transduction (Servet, Conde, Silva, & Zhou, 2010). For example, salt stress-induced histone-H3 levels via acetylation in maize plants ultimately balanced the cell wall proteins (Li et al., 2014). Lysine acetylation is a reversible modification of protein which is catalyzed by lysine acetyltransferase and deacetylase enzymatic activities (Friso & van Wijk, 2015). This can be considered as an epigenetic modification and extensively studied in plants as well as in other organisms (Berr, Shafiq, & Shen, 2011). Previous studies suggested that another type of histone lysine acetylation plays a major role in plant tolerance response to different abiotic and biotic stresses (Servet et al., 2010; Yuan, Liu, Luo, Yang, & Wu, 2013). Non-histone acetylation is widespread in mitochondria and chloroplast; however, lysine acetylation occurs at basic pH in the chloroplast stroma and mitochondrial matrix (Friso & van Wijk, 2015). In chloroplasts, several proteins such as Calvin cycle enzymes such as RuBisCO and some membrane proteins were found acetylated (Finkemeier, Laxa, Miguet, Howden, & Sweetlove, 2011). In mitochondria, the tricarboxylic acid cycle and some mitochondrial electron transport chain proteins are lysine-acetylated (König, Hartl, Boersema, Mann, & Finkemeier, 2014; Papanicolaou, O'Rourke, & Foster, 2014). König et al. (2014) identified 243 distinct acetylation sites at 120 lysine-acetylated mitochondrial proteins involved in the tricarboxylic acid cycle and protein metabolism. Mitochondrial acetylated proteins are also associated with many pathways such as histone/chromatin gene expression, signal transduction, and protein turnover (Anderson & Hirschey, 2012).

7.5 Conclusion

Proteomics has progressed at a rapid pace and a number of methods are now available for mapping the plant proteomes and identification of stress-responsive proteins with a high degree of accuracy and reproducibility. The efficacy of the majority of these methods has already been tested on the crop species; however, some of the recently developed methods such as AP-SWATH still needs to be fully utilized for the identification of protein candidates from the crops. Since working on agricultural crops needs more time because of their specific growing seasons, the information generated on the model plants can be translated on the agricultural crops to further understand their biology in greater detail. Moreover, PTM analysis using high-throughput approaches can further lead to the understanding of supraregulation of proteins in response to any external or internal factors and how their PTMs affect their functions and interaction with other proteins. This global analysis of proteins and their modifications will deepen our understanding of how crop functions and which kind of proteins needs to be targeted for the generation of improved cultivars with higher yield and stress tolerance.

References

- Abdallah, C., Dumas-Gaudot, E., Renaut, J., & Sergeant, K. (2012). Gel-based and gel-free quantitative proteomics approaches at a glance. *International Journal of Plant Genomics*, 2012, 494572.
- Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422, 198–207.
- Agrawal, G. K., Jwa, N. S., Iwahashi, Y., Yonekura, M., Iwahashi, H., & Rakwal, R. (2006). Rejuvenating rice proteomics: Facts, challenges, and visions. *Proteomics*, 6(20), 5549–5576.
- Agrawal, G. K., Jwa, N. S., & Rakwal, R. (2009). Rice proteomics: Ending phase I and the beginning of phase II. *Proteomics*, 9(4), 935–963.
- Agrawal, G. K., Pedreschi, R., Barkla, B. J., Bindschedler, L. V., Cramer, R., Sarkar, A., ... Rakwal, R. (2012). Translational plant proteomics: A perspective. *Journal of Proteomics*, 75(15), 4588–4601.
- Agrawal, G. K., & Rakwal, R. (2006). Rice proteomics: A cornerstone for cereal food crop proteomes. *Mass Spectrometry Reviews*, 25(1), 1–53.
- Agrawal, G. K., & Rakwal, R. (2008). *Plant proteomics: Technologies, strategies, and applications*. John Wiley & Sons. (Vol. 28).
- Agrawal, G. K., & Rakwal, R. (2011). Rice proteomics: A move toward expanded proteome coverage to comparative and functional proteomics uncovers the mysteries of rice and plant biology. *Proteomics*, 11(9), 1630–1649.
- Agrawal, G. K., Sarkar, A., Righetti, P. G., Pedreschi, R., Carpentier, S., Wang, T., ... Rampitsch, C. (2013). A decade of plant proteomics and mass spectrometry: Translation of technical advancements to food security and safety issues. *Mass Spectrometry Reviews*, 32, 335–365.
- Anderson, K. A., & Hirschey, M. D. (2012). Mitochondrial protein acetylation regulates metabolism. *Essays in Biochemistry*, 52, 23–35.
- Anguraj Vadivel, A. K. (2015). Gel-based proteomics in plants: time to move on from the tradition. *Frontiers in Plant Science*, 6, 369.
- Apweiler, R., Hermjakob, H., & Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta*, 1473, 4–8.
- Bateman, A., et al. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45, D158–D169.
- Berr, A., Shafiq, S., & Shen, W. H. (2011). Histone modifications in transcriptional activation during plant development. *Biochimica et Biophysica Acta*, 1809, 567–576.

- Bienvenu, W. V., Sumpton, D., Martinez, A., Lilla, S., Espagne, C., Meinel, T., & Giglione, C. (2012). Comparative large scale characterization of plant vs mammal proteins reveals similar and idiosyncratic N- α -acetylation features. *Molecular & Cellular Proteomics: MCP*, 11(6).
- Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Westermeier, R., & Postel, W. (1982). Isoelectric focusing in immobilized pH gradients: Principle, methodology and some applications. *Journal of Biochemical and Biophysical Methods*, 6, 317–339.
- Blasi, É. A., Buffon, G., Rativa, A. G., Lopes, M. C., Berger, M., Santi, L., ... Sperotto, R. A. (2017). High infestation levels of *Schizotetranychus oryzae* severely affects rice metabolism. *Journal of Plant Physiology*, 219, 100–111.
- Bridges, S. M., Magee, G. B., Wang, N., Williams, W. P., Burgess, S. C., & Nanduri, B. (2007). ProtQuant: A tool for the label-free quantification of MudPIT proteomics data. *BMC Bioinformatics*, 8(7), 1–9.
- Buffon, G., Blasi, E. A., Adamski, J. M., Ferla, N. J., Berger, M., Santi, L., ... Sperotto, R. A. (2016). Physiological and molecular alterations promoted by *Schizotetranychus oryzae* mite infestation in rice leaves. *Journal of Proteome Research*, 15(2), 431–446.
- Champagne, A., & Boutry, M. (2013). Proteomics of nonmodel plant species. *Proteomics*, 13(3–4), 663–673.
- Champion, A., Kreis, M., Mockaitis, K., Picaud, A., & Henry, Y. (2004). Arabidopsis kinome: After the casting. *Functional & Integrative Genomics*, 4, 163–187.
- Chen, X., Deng, Z., Yu, C., Yan, C., & Chen, J. (2016). Secretome analysis of rice suspension-cultured cells infected by *Xanthomonas oryzae* pv. *oryza* (Xoo). *Proteome Science*, 14, 2.
- Chen, X., Wei, S., Ji, Y., Guo, X., & Yang, F. (2015). Quantitative proteomics using SILAC: Principles, applications, and developments. *Proteomics*, 15(18), 3175–3192.
- Chevalier, F. (2010). Highlights on the capacities of "gel-based" proteomics. *Proteome Science*, 8(1), 23.
- Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E., & Mann, M. (2014). The growing landscape of lysine acetylation links metabolism and cell signalling. *Nature Reviews. Molecular Cell Biology*, 15, 536–550.
- Ciesla, J., Fraczyk, T., & Rode, W. (2011). Phosphorylation of basic amino acid residues in proteins: important but easily missed. *Acta Biochimica Polonica*, 58, 137–148.
- Cruz, E. R., Nguyen, H., Nguyen, T., & Wallace, I. S. (2019). Functional analysis tools for post-translational modification: A post-translational modification database for analysis of proteins and metabolic pathways. *The Plant Journal: for Cell and Molecular Biology*, 99(5), 1003–1013.
- Dai Vu, L., Gevaert, K., & De Smet, I. (2018). Protein language: Post-translational modifications talking to each other. *Trends in Plant Science*, 23(12), 1068–1080.
- Duan, G., & Walther, D. (2015). The roles of post-translational modifications in the context of protein interaction networks. *PLoS Computational Biology*, 11(2), e1004049.
- Evans, C., Noirel, J., Ow, S. Y., Salim, M., Pereira-Medrano, A. G., Couto, N., ... Zou, X. (2012). An insight into iTRAQ: where do we stand now? *Analytical and Bioanalytical Chemistry*, 404(4), 1011–1027.
- Everley, P. A., Krijgsveld, J., Zetter, B. R., & Gygi, S. P. (2004). Quantitative cancer proteomics: stable isotope labeling with amino acids in cell culture (SILAC) as a tool for prostate cancer research. *Molecular & Cellular Proteomics: MCP*, 3, 729–735.
- Ferrández-Ayela, A., Micol-Ponce, R., Sánchez-García, A. B., Alonso-Peral, M. M., Micol, J. L., & Ponce, M. R. (2013). Mutation of an Arabidopsis NatB N-alpha-terminal acetylation complex component causes pleiotropic developmental defects. *PLoS One*, 8, e80697.
- Finkemeier, I., Laxa, M., Miguet, L., Howden, A. J., & Sweetlove, L. J. (2011). Proteins of diverse function and subcellular location are lysine acetylated in Arabidopsis. *Plant Physiology*, 155, 1779–1790.
- Fitchette, A., Cabanes-Macheteau, M., Marvin, L., Martin, B., Satiat-Jeunemaitre, B., Gomord, V., ... Hawes, C. (1999). Biosynthesis and immunolocalization of Lewis a-containing N-glycans in the plant cell. *Plant Physiology*, 121, 333–344.
- Forment, J., Naranjo, M. Á., Roldán, M., Serrano, R., & Vicente, O. (2002). Expression of Arabidopsis SR-like splicing proteins confers salt tolerance to yeast and transgenic plants. *The Plant Journal: for Cell and Molecular Biology*, 30(5), 511–519.
- Foster, L. J., De Hoog, C. L., & Mann, M. (2003). Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 5813–5818.
- Friso, G., & Van Wijk, K. J. (2015). Posttranslational protein modifications in plant metabolism. *Plant Physiology*, 169, 1469–1487.
- García, A., Senis, Y., Tomlinson, M. G., & Watson, S. P. (2007). Platelet genomics and proteomics. *Platelets*, 99–116.
- Ge, P., Hao, P., Cao, M., Guo, G., Lv, D., Subburaj, S., ... Yan, Y. (2013). iTRAQ-based quantitative proteomic analysis reveals new metabolic pathways of wheat seedling growth under hydrogen peroxide stress. *Proteomics*, 13(20), 3046–3058.
- Ghatak, A., Chaturvedi, P., Paul, P., Agrawal, G. K., Rakwal, R., Kim, S. T., ... Gupta, R. (2017). Proteomics survey of Solanaceae family: Current status and challenges ahead. *Journal of Proteomics*, 169, 41–57.
- Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., ... Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics: MCP*, 11(6). Available from <https://doi.org/10.1074/mcp.O111.016717>.
- Gnad, F., Gunawardena, J., & Mann, M. (2010). PHOSIDA 2011: The posttranslational modification database. *Nucleic Acids Research*, 39, D253–D260.
- Gomord, V., & Faye, L. (2004). Posttranslational modification of therapeutic proteins in plants. *Current Opinion in Plant Biology*, 7(2), 171–181.
- Gomord, V., Fitchette, A. C., Menu-Bouaouiche, L., Saint-Jore-Dupas, C., Plasson, C., Michaud, D., & Faye, L. (2010). Plant-specific glycosylation patterns in the context of therapeutic protein production. *Plant Biotechnology Journal*, 8(5), 564–587.
- Gupta, R., Lee, S. J., Min, C. W., Kim, S. W., Park, K. H., Bae, D. W., ... Kim, S. T. (2016a). Coupling of gel-based 2-DE and 1-DE shotgun proteomics approaches to dig deep into the leaf senescence proteome of *Glycine max*. *Journal of Proteomics*, 148, 65–74.
- Gupta, R., Lee, S. J., Min, C. W., Kim, S. W., Park, K. H., Bae, D. W., ... Kim, S. T. (2016b). Proteome data associated with the leaf senescence in *Glycine max*. *Data Br*, 9, 90–95.

- Gupta, R., Min, C. W., Kim, S. W., Wang, Y., Agrawal, G. K., Rakwal, R., . . . Bae, D. W. (2015). Comparative investigation of seed coats of brown- vs yellow-colored soybean seeds using an integrated proteomics and metabolomics approach. *Proteomics*, *15*, 1706–1716.
- Gupta, R., Min, C. W., Kim, S. W., Yoo, J. S., Moon, A. R., Shin, A. Y., . . . Kim, S. T. (2020). A TMT-based quantitative proteome analysis to elucidate the TSWV induced signaling cascade in susceptible and resistant cultivars of *Solanum lycopersicum*. *Plants*, *9*(3), 290.
- Gupta, R., Min, C. W., Kim, Y. J., & Kim, S. T. (2019). Identification of Msp1-induced signaling components in rice leaves by integrated proteomic and phosphoproteomic analysis. *Int. J. Mol. Sci.*, *20*.
- Gupta, R., Min, C. W., Kramer, K., Agrawal, G. K., Rakwal, R., Park, K. H., . . . Kim, S. T. (2018). A multi-omics analysis of *Glycine max* leaves reveals alteration in flavonoid and isoflavonoid metabolism upon ethylene and abscisic acid treatment. *Proteomics*, *18*(7), e1700366.
- Gupta, R., Min, C. W., Meng, Q., Agrawal, G. K., Rakwal, R., & Kim, S. T. (2018). Comparative phosphoproteome analysis upon ethylene and abscisic acid treatment in *Glycine max* leaves. *Plant Physiology and Biochemistry: PPB/Societe Francaise de Physiologie Vegetale*, *130*, 173–180.
- Gupta, R., Min, C. W., Meng, Q., Jun, T. H., Agrawal, G. K., Rakwal, R., & Kim, S. T. (2018). Phosphoproteome data from abscisic acid and ethylene treated *Glycine max* leaves. *Data in Brief*, *20*, 516–520.
- Gupta, R., Wang, Y., Agrawal, G. K., Rakwal, R., Jo, I. H., Bang, K. H., & Kim, S. T. (2015). Time to dig deep into the plant proteome: A hunt for low-abundance proteins. *Frontiers in Plant Science*, *6*, 1–3.
- Harmoko, R., Yoo, J. Y., Ko, K. S., Ramasamy, N. K., Hwang, B. Y., Lee, E. J., . . . Lee, S. (2016). N-glycan containing a core α 1,3-fucose residue is required for basipetal auxin transport and gravitropic response in rice (*Oryza sativa*). *The New Phytologist*, *212*, 108–122.
- Hashiguchi, A., & Komatsu, S. (2016). Impact of post-translational modifications of crop proteins under abiotic stress. *Proteomes*, *4*(4), 42.
- Hayashi, G., Moro, C. F., Rohila, J. S., Shibato, J., Kubo, A., Imanaka, T., . . . Ichikawa, K. (2015). 2D-DIGE-based proteome expression changes in leaves of rice seedlings exposed to low-level gamma radiation at litate village, Fukushima. *Plant Signaling & Behavior*, *10*(12), e1103406.
- Hebert, D. N., Lamriben, L., Powers, E. T., & Kelly, J. W. (2014). The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis. *Nature Chemical Biology*, *10*, 902–910.
- Hou, H., Wang, P., Zhang, H., Wen, H., Gao, F., Ma, N., . . . Li, L. (2015). Histone acetylation is involved in gibberellin-regulated sodCp gene expression in maize aleurone layers. *Plant & Cell Physiology*, *56*(11), 2139–2149.
- Hu, Y., Qin, F., Huang, L., Sun, Q., Li, C., Zhao, Y., & Zhou, D. X. (2009). Rice histone deacetylase genes display specific expression patterns and developmental functions. *Biochemical and Biophysical Research Communications*, *388*(2), 266–271.
- Huang, H., Arighi, C. N., Ross, K. E., Ren, J., Li, G., Chen, S. C., . . . Wu, C. H. (2017). iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Research*, *46*, D542–D550.
- Jeffers, V., & Sullivan, W. J. (2012). Lysine acetylation is widespread on proteins of diverse function and localization in the protozoan parasite *Toxoplasma gondii*. *Eukaryotic Cell*, *11*, 735–742.
- Jia, Y., Liu, H., Qu, Z., Wang, J., Wang, X., Wang, Z., . . . Zhao, H. (2020). Transcriptome sequencing and iTRAQ of different rice cultivars provide insight into molecular mechanisms of cold-tolerance response in japonica rice. *Rice*, *13*, 43.
- Jones, A. M., Bennett, M. H., Mansfield, J. W., & Grant, M. (2006). Analysis of the defence phosphoproteome of *Arabidopsis thaliana* using differential mass tagging. *Proteomics*, *6*, 4155–4165.
- Jorin-Novo, J. V., Maldonado, A. M., Echevarria-Zomeno, S., Valledor, L., Castillejo, M. A., Curto, M., . . . Redondo, I. (2009). Plant proteomics update (2007–2008): second-generation proteomic techniques, an appropriate experimental design, and data analysis to fulfill MIAPE standards, increase plant proteome coverage and expand biological knowledge. *Journal of Proteomics*, *72*(3), 285–314.
- Kant, P., Kant, S., Gordon, M., Shaked, R., & Barak, S. (2007). Stress Response Suppressor1 and Stress Response Suppressor2, two Dead-box RNA helicases that attenuate Arabidopsis responses to multiple abiotic stresses. *Plant Physiology*, *145*(3), 814–830.
- Kao, H. J., Jhong, J. H., Huang, K. Y., Cheng, K. H., Su, M. G., Hsieh, Y. C., . . . Huang, H. D. (2015). dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Research*, *44*(D1), D435–D446.
- Kim, J. Y., Park, S. J., Jang, B., Jung, C. H., Ahn, S. J., Goh, C. H., . . . Kang, H. (2007). Functional characterization of a glycine-rich RNA-binding protein 2 in *Arabidopsis thaliana* under abiotic stress conditions. *The Plant Journal: for Cell and Molecular Biology*, *50*(3), 439–451.
- Kim, S. T., Kim, S. G., Agrawal, G. K., Kikuchi, S., & Rakwal, R. (2014). Rice proteomics: a model system for crop improvement and food security. *Proteomics*, *14*, 593–610.
- Kim, Y. O., Kim, J. S., & Kang, H. (2005). Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in *Arabidopsis thaliana*. *The Plant Journal: for Cell and Molecular Biology*, *42*(6), 890–900.
- Klose, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. *Humangenetik*, *26*(3), 231–243.
- Ko, J. H., Yang, S. H., & Han, K. H. (2006). Upregulation of an Arabidopsis RING-H2 gene, XERICO, confers drought tolerance through increased abscisic acid biosynthesis. *The Plant Journal: for Cell and Molecular Biology*, *47*(3), 343–355.
- Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L., Deciu, C., Haynes, P. A., . . . Wolters, D. (2002). Proteomic survey of metabolic pathways in rice. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(18), 11969–11974.
- Komatsu, S., Kamal, A. H. M., & Hossain, Z. (2014). Wheat proteomics: Proteome modulation and abiotic stress acclimation. *Front. Plant Sci.*, *5*, 684.
- König, A. C., Hartl, M., Boersema, P. J., Mann, M., & Finkemeier, I. (2014). The mitochondrial lysine acetylome of *Arabidopsis*. *Mitochondrion*, *19*, 252–260.
- Kumar, R., Kumar, A., Subba, P., Gayali, S., Barua, P., Chakraborty, S., & Chakraborty, N. (2014). Nuclear phosphoproteome of developing chickpea seedlings (*Cicer arietinum* L.) and protein-kinase interaction network. *Journal of Proteomics*, *105*, 58–73.
- Kwon, S. J., Choi, E. Y., Choi, Y. J., Ahn, J. H., & Park, O. K. (2006). Proteomics studies of post-translational modifications in plants. *Journal of Experimental Botany*, *57*(7), 1547–1551.

- Laemmli, U. K. (1971). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature (Lond.)*, 227, 680–685.
- Lecomte, P., Urbach, S., Demetree, E., Biron, D. G., Langin, T., & Bonhomme, L. (2014). Using SILAC strategy to identify protein effectors in the wheat-*Fusarium graminearum* pathosystem. *Effectome Network*.
- Li, H., Yan, S., Zhao, L., Tan, J., Zhang, Q., Gao, F., ... Li, L. (2014). Histone acetylation associated up-regulation of the cell wall related genes is involved in salt stress induced maize root swelling. *BMC Plant Biology*, 14(1), 105.
- Li, M., Chen, X., He, D., & Yang, P. (2020). Proteomic analysis reveals that calcium channel blockers affect radicle protrusion during rice seed germination. *Plant Growth Regulation*, 90(2), 393–407.
- Linster, E., Stephan, I., Bienvenut, W. V., Maple-Grødem, J., Myklebust, L. M., Huber, M., ... Arnesen, T. (2015). Downregulation of N-terminal acetylation triggers ABA-mediated drought responses in Arabidopsis. *Nature Communications*, 6, 7640.
- Liu, Y., Cao, D., Ma, L., Jin, X., Yang, P., Ye, F., ... Wei, C. (2018). TMT-based quantitative proteomics analysis reveals the response of tea plant (*Camellia sinensis*) to fluoride. *Journal of Proteomics*, 176, 71–81.
- Liu, Y., Lu, S., Liu, K., Wang, S., Huang, L., & Guo, L. (2019). Proteomics: A powerful tool to study plant responses to biotic stress. *Plant Methods*, 15, 135.
- Lv, D. W., Zhu, G. R., Zhu, D., Bian, Y. W., Liang, X. N., Cheng, Z. W., ... Yan, Y. M. (2016). Proteomic and phosphoproteomic analysis reveals the response and defense mechanism in leaves of diploid wheat *T. monococcum* under salt stress and recovery. *Journal of Proteomics*, 143, 93–105.
- Ma, Q., Shi, C., Su, C., & Liu, Y. (2020). Complementary analyses of the transcriptome and iTRAQ proteome revealed mechanism of ethylene dependent salt response in bread wheat (*Triticum aestivum* L.). *Food Chemistry*, 126866.
- Maclean, J. L., Dawe, D. C., Hardy, B., & Hettel, G. P. (2002). *Rice almanac* (3rd ed.). International Rice Research Institute, Los Baños, 253.
- Maor, R., Jones, A., Nühse, T. S., Studholme, D. J., Peck, S. C., & Shirasu, K. (2007). Multidimensional protein identification technology (MudPIT) analysis of ubiquitinated proteins in plants. *Molecular & Cellular Proteomics: MCP*, 6(4), 601–610.
- Meng, Q., Gupta, R., Min, C. W., Kim, J., Kramer, K., Wang, Y., ... Kim, S. T. (2018). Label-free quantitative proteome data associated with MSP1 and flg22 induced signaling in rice leaves. *Data Br*, 20, 204–209.
- Meng, Q., Gupta, R., Min, C. W., Kwon, S. W., Wang, Y., Je, B. I., ... Kim, S. T. (2019). Proteomics of rice—*Magnaporthe oryzae* interaction: What have we learned so far? *Frontiers in Plant Science*, 10.
- Min, C. W., Gupta, R., Agrawal, G. K., Rakwal, R., & Kim, S. T. (2019). Concepts and strategies of soybean seed proteomics using the shotgun proteomics approach. *Expert Review of Proteomics*, 16, 795–804.
- Min, C. W., Park, J., Bae, J. W., Agrawal, G. K., Rakwal, R., Kim, Y., ... Gupta, R. (2020). In-depth investigation of low-abundance proteins in matured and filling stages seeds of *Glycine max* employing a combination of protamine sulfate precipitation and TMT-based quantitative proteomic analysis. *Cells*, 9(6), 1517.
- Mithoe, S. C., & Menke, F. L. (2011). Phosphoproteomics perspective on plant signal transduction and tyrosine phosphorylation. *Phytochemistry*, 72, 997–1006.
- Moremen, K. W., Tiemeyer, M., & Nairn, A. V. (2012). Vertebrate protein glycosylation: Diversity, synthesis and function. *Nature Reviews. Molecular Cell Biology*, 13, 448–462.
- Nallamilli, B. R., Edelmann, M. J., Zhong, X., Tan, F., Mujahid, H., Zhang, J., ... Peng, Z. (2014). Global analysis of lysine acetylation suggests the involvement of protein acetylation in diverse biological processes in rice (*Oryza sativa*). *PLoS One*, 9, e89283.
- Nühse, T. S., Bottrill, A. R., Jones, A. M., & Peck, S. C. (2007). Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses. *The Plant Journal: for Cell and Molecular Biology*, 51, 931–940.
- O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *The Journal of Biological Chemistry*, 250, 4007–4021.
- Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics: MCP*, 1, 376–386.
- Ong, S. E., Mittler, G., & Mann, M. (2004). Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nature Methods*, 1, 119–126.
- Pandey, R., Muller, A., Napoli, C. A., Selinger, D. A., Pikaard, C. S., Richards, E. J., ... Jorgensen, R. A. (2002). Analysis of histone acetyltransferase and histone deacetylase families of *Arabidopsis thaliana* suggests functional diversification of chromatin modification among multicellular eukaryotes. *Nucleic Acids Research*, 30, 5036–5055.
- Papanicolaou, K. N., O'Rourke, B., & Foster, D. B. (2014). Metabolism leaves its mark on the powerhouse: Recent progress in post-translational modifications of lysine in mitochondria. *Frontiers in Physiology*, 5, 301.
- Pechanova, O., Takáč, T., Šamaj, J., & Pechan, T. (2013). Maize proteomics: An insight into the biology of an important cereal crop. *Proteomics*, 13, 637–662.
- Perham, R. N. (2000). Swinging arms and swinging domains in multifunctional enzymes: Catalytic machines for multistep reactions. *Annual Review of Biochemistry*, 69, 961–1004.
- Pesaresi, P., Gardner, N. A., Masiero, S., Dietzmann, A., Eichacker, L., Wickner, R., ... Leister, D. (2003). Cytoplasmic N-terminal protein acetylation is required for efficient photosynthesis in Arabidopsis. *The Plant Cell*, 15, 1817–1832.
- Rakwal, R., & Agrawal, G. K. (2003). Rice proteomics: Current status and future perspectives. *Electrophoresis*, 24(19–20), 3378–3389.
- Rakwal, R., Hayashi, G., Shibato, J., Deepak, S. A., Gundimeda, S., Simha, U., ... Kubo, A. (2018). Progress toward rice seed OMICS in low-level gamma radiation environment in Iitate Village, Fukushima. *The Journal of Heredity*, 109, 206–211.
- Rampitsch, C., & Huang, M. (2019). Temporal quantitative changes in the resistant and susceptible wheat leaf apoplastic proteome during infection by wheat leaf rust (*Puccinia triticina*). *Frontiers in Plant Science*, 10, 1291.
- Righetti, P. G. (2014). The Monkey King: A personal view of the long journey towards a proteomic Nirvana. *Journal of Proteomics*, 107, 39–49.

- Riter, L. S., Jensen, P. K., Ballam, J. M., Urbanczyk-Wochniak, E., Clough, T., Vitek, O., ... MacIsaac, S. (2011). Evaluation of label-free quantitative proteomics in a plant matrix: A case study of the night-to-day transition in corn leaf. *Analytical Methods*, 3(12), 2733–2739.
- Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., ... Purkayastha, S. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics: MCP*, 3(12), 1154–1169.
- Roussel, M., Peltier, J. B., Mock, H. P., Matros, A., Maldonado, A. M., & Jorrín, J. V. (2006). Plant proteome analysis: a 2004–2006 update. *Proteomics*, 6(20), 5529–5548.
- Sadowski, I., Breitkreutz, B. J., Stark, C., Su, T. C., Dahabieh, M., Raithatha, S., Bernhard, W., Oughtred, R., Dolinski, K., Barreto, K., & Tyers, M. (2013). The PhosphoGRID *Saccharomyces cerevisiae* protein phosphorylation site database: Version 2.0 update. *Database*, 2013, bat026.
- Sarkar, A., Islam, M. T., Zargar, S. M., Dogra, V., Kim, S. T., Gupta, R., ... Sirisaththa, S. (2014). Proteomics potential and its contribution toward sustainable agriculture. *Agroecology, Ecosystems, and Sustainability*, 20, 151.
- Scheele, G. A. (1975). Two-dimensional gel analysis of soluble proteins. Characterization of guinea pig exocrine pancreatic proteins. *Journal of Biological Chemistry*, 250, 5375–5385.
- Schlösser, A., & Lehmann, W. D. (2000). Five-membered ring formation in unimolecular reactions of peptides: a key structural element controlling low-energy collision-induced dissociation of peptides. *Journal of Mass Spectrometry*, 35(12), 1382–1390.
- Schönberg, A., & Baginsky, S. (2012). Signal integration by chloroplast phosphorylation networks: An update. *Frontiers in Plant Science*, 3, 256.
- Servet, C., Conde, E., Silva, N., & Zhou, D. X. (2010). Histone acetyltransferase AtGCN5/HAG1 is a versatile regulator of developmental and inducible gene expression in Arabidopsis. *Molecular Plant*, 3, 670–677.
- Sew, Y. S., Aizat, W. M., Razak, M. S. F. A., Zainal-Abidin, R. A., Simoh, S., & Abu-Bakar, N. (2020). Comprehensive proteomics data on whole rice grain of selected pigmented and non-pigmented rice varieties using SWATH-MS approach. *Data in Brief*, 105927.
- Silva-Sanchez, C., Li, H., & Chen, S. (2015). Recent advances and challenges in plant phosphoproteomics. *Proteomics*, 15, 1127–1141.
- Singh, A., Giri, J., Kapoor, S., Tyagi, A. K., & Pandey, G. K. (2014). Protein phosphatase complement in rice: Genome-wide identification and transcriptional analysis under abiotic stress conditions and reproductive development. *BMC Genom*, 15.
- Smith, L. M., Kelleher, N. L., Linial, M., Goodlett, D., Langridge-Smith, P., Goo, Y. A., ... Rontree, J. (2013). Proteoform: A single term describing protein complexity. *Nature Methods*, 10(3), 186–187.
- Song, W., Mentink, R. A., Henquet, M. G., Cordewener, J. H., van Dijk, A. D., Bosch, D., ... van der Krol, A. R. (2013). N-Glycan occupancy of Arabidopsis N-glycoproteins. *Journal of Proteomics*, 93, 343–355.
- Song, Y., Zhang, C., Ge, W., Zhang, Y., Burlingame, A. L., & Guo, Y. (2011). Identification of NaCl stress-responsive apoplastic proteins in rice shoot stems by 2D-DIGE. *Journal of Proteomics*, 74, 1045–1067.
- Spoel, S. H. (2018). Orchestrating the proteome with post-translational modifications. *Journal of Experimental Botany*, 69(19), 4499–4503.
- Strasser, R., Altmann, F., Mach, L., Glössl, J., & Steinkellner, H. (2004). Generation of *Arabidopsis thaliana* plants with complex N-glycans lacking β 1, 2-linked xylose and core α 1, 3-linked fucose. *FEBS Letters*, 561(1–3), 132–136.
- Sunkar, R., Kapoor, A., & Zhu, J. K. (2006). Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. *The Plant Cell*, 18(8), 2051–2065.
- Takano, S., Matsuda, S., Funabiki, A., Furukawa, J., Yamauchi, T., Tokuji, Y., ... Kato, K. (2015). The rice RCN11 gene encodes β 1,2 xylosyltransferase and is required for plant responses to abiotic stresses and phytohormones. *Plant Science (Shannon, Ireland)*, 236, 75–88.
- Teshima, R., Nakamura, R., Satoh, R., & Nakamura, R. (2010). 2D-DIGE analysis of rice proteins from different cultivars. *Regulatory Toxicology and Pharmacology: RTP*, 58, S30–S35.
- Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., ... Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8), 1895–1904.
- Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., & Nesvizhskii, A. I. (2015). DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods*, 12(3), 258–264.
- Unlu, M., Morgan, M. E., & Minden, J. S. (1997). Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis*, 18, 2071–2077.
- van Bentem, S. D. L. F., Anrather, D., Roitinger, E., Djamei, A., Hufnagl, T., Barta, A., ... Hirt, H. (2006). Phosphoproteomics reveals extensive in vivo phosphorylation of Arabidopsis proteins involved in RNA metabolism. *Nucleic Acids Research*, 34, 3267–3278.
- van Wijk, K. J., Friso, G., Walther, D., & Schulze, W. X. (2014). Meta-analysis of *Arabidopsis thaliana* phospho-proteomics data reveals compartmentalization of phosphorylation motifs. *The Plant Cell*, 26, 2367–2389.
- Vandamme, J., Castermans, D., & Thevelein, J. M. (2012). Molecular mechanisms of feedback inhibition of protein kinase A on intracellular cAMP accumulation. *Cellular Signalling*, 24(8), 1610–1618.
- Vanderschuren, H., Lentz, E., Zainuddin, I., & Gruissem, W. (2013). Proteomics of model and crop plant species: Status, current limitations and strategic advances for crop improvement. *Journal of Proteomics*, 93, 5–19.
- Via, A., Gould, C. M., Chica, C., Dinkel, H., Jensen, L. J., Diella, F., & Gibson, T. J. (2010). Phospho. ELM: A database of phosphorylation sites—update 2011. *Nucleic Acids Research*, 39, D261–D267.
- Viñtor, R., Loutelier-Bourhis, C., Fichette, A., Margerie, P., Gonneau, M., Faye, L., & Lerouge, P. (2003). Protein N-glycosylation is similar in the moss *Physcomitrella patens* and in higher plants. *Planta*, 218, 269–275.
- Villafañez, F., Gottifredi, V., & Soria, G. (2019). Development and optimization of a miniaturized western blot-based screening platform to identify regulators of post-translational modifications. *High-throughput*, 8(2), 15.

- von Schaewen, A., Sturm, A., O'Neill, J., & Chrispeels, M. (1993). Isolation of a mutant *Arabidopsis* plant that lacks N-acetyl glucosaminyl transferase I and is unable to synthesize Golgi-modified complex N-linked glycans. *Plant Physiology*, *102*, 1109–1118.
- Walsh, G., & Jefferis, R. (2006). Post-translational modifications in the context of therapeutic proteins. *Nature Biotechnology*, *24*(10), 1241–1252.
- Wang, Y., Liu, Z., Cheng, H., Gao, T., Pan, Z., Yang, Q., . . . Xue, Y. (2014). EKPD: A hierarchical database of eukaryotic protein kinases and protein phosphatases. *Nucleic Acids Research*, *42*, D496–D502.
- Wang, B., Ren, Y., Lu, C., & Wang, X. (2015). iTRAQ-based quantitative proteomics analysis of rice leaves infected by rice stripe virus reveals several proteins involved in symptom formation. *Virology Journal*, *12*(1), 99.
- Wang, G., Li, H., Wang, K., Yang, J., Duan, M., Zhang, J., & Ye, N. (2020). Regulation of gene expression involved in the remobilization of rice straw carbon reserves results from moderate soil drying during grain filling. *The Plant Journal: For Cell and Molecular Biology*, *101*(3), 604–618.
- Wang, M., Yue, H., Feng, K., Deng, P., Song, W., & Nie, X. (2016). Genome-wide identification, phylogeny and expression profiles of mitogen activated protein kinase kinase kinase (MAPKKK) gene family in bread wheat (*Triticum aestivum* L.). *BMC Genomics*, *17*(1), 668.
- Wang, Y., Gupta, R., Song, W., Huh, H. H., Lee, S. E., Wu, J., . . . Kim, S. T. (2017). Label-free quantitative secretome analysis of *Xanthomonas oryzae* pv. *oryzae* highlights the involvement of a novel cysteine protease in its pathogenicity. *Journal of Proteomics*, *169*, 202–214.
- Washburn, M. P., Wolters, D., & Yates, J. R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, *19*(3), 242–247.
- Webster, D. E., & Thomas, M. C. (2012). Post-translational modification of plant-made foreign proteins; glycosylation and beyond. *Biotechnology Advances*, *30*(2), 410–418.
- Wei, K., & Pan, S. (2014). Maize protein phosphatase gene family: Identification and molecular characterization. *BMC Genomics*, *15*(1), 773.
- Wiese, S., Reidegeld, K. A., Meyer, H. E., & Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics*, *7*(3), 340–350.
- Williams, D. B. (2006). Beyond lectins: The calnexin/calreticulin chaperone system of the endoplasmic reticulum. *Journal of Cell Science*, *119*, 615–623.
- Wilson, I., Zeleny, R., Kolarich, D., Staudacher, E., Stroop, C., Kamlering, J., & Altmann, F. (2001). Analysis of Asn-linked glycans from vegetable foodstuffs: Widespread occurrence of Lewis a, core alpha1,3-linked fucose and xylose substitutions. *Glycobiology*, *11*, 261–274.
- Wong, C. H. (2005). Protein glycosylation: new challenges and opportunities. *The Journal of Organic Chemistry*, *70*, 4219–4225.
- Wray, W., Boulikas, T., Wray, V. P., & Hancock, R. (1981). Silver staining of proteins in polyacrylamide gels. *Analytical Biochemistry*, *118*, 197–203.
- Wu, Q., Zhang, C., Chen, Y., Zhou, K., Zhan, Y., & Jiang, D. (2020). OsCpn60β1 is essential for chloroplast development in rice (*Oryza sativa* L.). *Int. J. Mol. Sci.*, *21*(11), 4023.
- Wu, X., Sklodowski, K., Encke, B., & Schulze, W. X. (2014). A kinase-phosphatase signaling module with BSK8 and BSL2 involved in regulation of sucrose-phosphate synthase. *Journal of Proteome Research*, *13*, 3397–3409.
- Wu, Y., Mirzaei, M., Pascovici, D., Haynes, P. A., & Atwell, B. J. (2019). Proteomes of leaf-growing zones in rice genotypes with contrasting drought tolerance. *Proteomics*, *19*(9), 1800310.
- Xiao, R., Li, L., & Ma, Y. (2019). A label-free proteomic approach differentiates between conventional and organic rice. *Journal of Food Composition and Analysis: An Official Publication of the United Nations University, International Network of Food Data Systems*, *80*, 51–61.
- Xing, T., Ouellet, T., & Miki, B. L. (2002). Towards genomic and proteomic studies of protein phosphorylation in plant–pathogen interactions. *Trends in Plant Science*, *7*(5), 224–230.
- Xu, F., Huang, Y., Li, L., Gannon, P., Linster, E., Huber, M., . . . Hell, R. (2015). Two N-terminal acetyltransferases antagonistically regulate the stability of a nod-like receptor in *Arabidopsis*. *The Plant Cell*, *27*(5), 1547–1562.
- Yang, P., Li, Y., He, C., Yan, J., Zhang, W., Li, X., . . . Liu, X. (2020). Phenotype and TMT-based quantitative proteomics analysis of *Brassica napus* reveals new insight into chlorophyll synthesis and chloroplast structure. *Journal of Proteomics*, *214*, 103621.
- Yao, Q., Ge, H., Wu, S., Zhang, N., Chen, W., Xu, C., . . . Xu, D. (2014). P3DB 3.0: from plant phosphorylation sites to protein networks. *Nucleic Acids Research*, *42*(D1), D1206–D1213.
- Yin, X., He, D., Gupta, R., & Yang, P. (2015). Physiological and proteomic analyses on artificially aged *Brassica napus* seed. *Frontiers in Plant Science*, *6*, 112.
- Ytterberg, A. J., & Jensen, O. N. (2010). Modification-specific proteomics in plant biology. *Journal of Proteomics*, *73*, 2249–2266.
- Yu, B., Li, J., Koh, J., Dufresne, C., Yang, N., Qi, S., . . . Li, H. (2016). Quantitative proteomics and phosphoproteomics of sugar beet monosomic addition line M14 in response to salt stress. *Journal of Proteomics*, *143*, 286–297.
- Yuan, L., Liu, X., Luo, M., Yang, S., & Wu, K. (2013). Involvement of histone modifications in plant abiotic stress responses. *Journal of Integrative Plant Biology*, *55*, 892–901.
- Zhang, F., Wang, L., Lim, J. Y., Kim, T., Pyo, Y., Sung, S., . . . Qiao, H. (2016). Phosphorylation of CBP20 links microRNA to root growth in the ethylene response. *PLoS Genetics*, *12*(11), e1006437.
- Zhang, H., He, D., Yu, J., Li, M., Damaris, R. N., Gupta, R., . . . Yang, P. (2016). Analysis of dynamic protein carbonylation in rice embryo during germination through AP-SWATH. *Proteomics*, *16*(6), 989–1000.
- Zhang, H., Zhou, H., Berke, L., Heck, A. J., Mohammed, S., Scheres, B., & Menke, F. L. (2013). Quantitative phosphoproteomics after auxin-stimulated lateral root induction identifies an SNX1 protein phosphorylation site required for growth. *Molecular & Cellular Proteomics: MCP*, *12*(5), 1158–1169.

- Zhang, M., Lv, D., Ge, P., Bian, Y., Chen, G., Zhu, G., . . . Yan, Y. (2014). Phosphoproteome analysis reveals new drought response and defense mechanisms of seedling leaves in bread wheat (*Triticum aestivum* L.). *Journal of Proteomics*, *109*, 290–308.
- Zhou, H., Wang, X., Huo, C., Wang, H., An, Z., Sun, D., . . . Zhang, B. (2019). A quantitative proteomics study of early heat-regulated proteins by two-dimensional difference gel electrophoresis identified OsUBP21 as a negative regulator of heat stress responses in rice. *Proteomics*, *19*(20), 1900153.
- Zhu, F. Y., Chen, M. X., Su, Y. W., Xu, X., Ye, N. H., Cao, Y. Y., . . . Jin, Y. (2016). SWATH-MS quantitative analysis of proteins in the rice inferior and superior spikelets during grain filling. *Frontiers in Plant Science*, *7*, 1926.
- Zhu, W., Smith, J. W., & Huang, C. M. (2010). Mass spectrometry-based label-free quantitative proteomics. *Journal of Biomedicine and Biotechnology*, *10*, 840518, 2010.
- Zielinska, D. F., Gnad, F., Schropp, K., Wisniewski, J. R., & Mann, M. (2012). Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. *Molecular Cell*, *46*, 542–548.
- Zörb, C., Schmitt, S., & Mühling, K. H. (2010). Proteomic changes in maize roots after short-term adjustment to saline growth conditions. *Proteomics*, *10*(24), 4441–4449.
- Zulawski, M., Braginets, R., & Schulze, W. X. (2013). PhosPhAt goes kinases: searchable protein kinase target information in the plant phosphorylation site database PhosPhAt. *Nucleic Acids Research*, *41*, D1176–D1184.

Further reading

- Kosová, K., Vítámvás, P., Prášil, I. T., & Renaut, J. (2011). Plant proteome changes under abiotic stress—contribution of proteomics studies to understanding plant stress response. *Journal of Proteomics*, *74*(8), 1301–1322.
- Olszewski, N. E., West, C. M., Sassi, S. O., & Hartweck, L. M. (2010). O-GlcNAc protein modification in plants: Evolution and function. *Biochimica et Biophysica Acta*, *1800*, 49–56.
- Strasser, R. (2016). Plant protein glycosylation. *Glycobiology*, *26*(9), 926–939.
- Tan, B. C., Lim, Y. S., & Lau, S. E. (2017). Proteomics in commercial crops: An overview. *Journal of Proteomics*, *169*, 176–188.
- Wang, H., Chevalier, D., Larue, C., Ki Cho, S., & Walker, J. C. (2007). The protein phosphatases and protein kinases of *Arabidopsis thaliana*. *Arabidopsis Book*, *5*, e0106.
- Xu, C., Wang, S., Thibault, G., & Ng, D. T. (2013). Futile protein folding cycles in the ER are terminated by the unfolded protein O-mannosylation pathway. *Science (New York, N.Y.)*, *340*, 978–981.
- Yan, J. X., Harry, R. A., Spibey, C., & Dunn, M. J. (2000). Postelectrophoretic staining of proteins separated by two-dimensional gel electrophoresis using SYPRO dyes. *Electrophoresis*, *21*(17), 3657–3665.
- Zieske, L. R. (2006). A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. *Journal of Experimental Botany*, *57*(7), 1501–1508.

Metabolomics and sustainable agriculture: concepts, applications, and perspectives

Noureddine Benkeblia

Department of Life Sciences, The Biotechnology Center, The University of the West Indies, Kingston, Jamaica

8.1 Introduction

Although wild crops have been collected and eaten from more than 100,000 years ago, domestication of plants and animals and, development and dissemination of agricultural techniques occurred later from around 9500 BCE. From the 19th century and the Industrial Revolution, the so-called “conventional agriculture” or “industrial agriculture” paved the way for new farming techniques and improved livestock breeding leading to greater food production. Beginning in the mid-20th century, the Green Revolution leads to a great increase in food production resulting from the introduction of new and high-yielding varieties, chemicals, and intensive mechanization. This modern agriculture that depends to large extent on synthetic fertilizers and pesticides was viewing the farm as a factory with inputs and outputs with the main goal is to prioritize higher yields. According to the World Bank, food production increased by 70%–90% in the past 50 years resulting in conventional agriculture rather than greater cultivated acreage. Unfortunately, these new varieties cropped intensively require large inputs including large amounts of fertilizers and pesticides, therefore raising major concerns about harmful and negative impacts on the environment.

The term “sustainable agriculture” is a complex concept with environmental, economic, and social facets and its definition is versatile. The concept of sustainable agriculture was first stated by Rodale (1988), and many other definitions have been suggested. These suggestions were analyzed by Okigbo (1991) who defined “agricultural sustainability as one that maintains an acceptable and increasing level of productivity that satisfies prevailing needs and is continuously adapted to meet the future needs for increasing the carrying capacity of the resource base and other worthwhile human needs,” and accordingly when resources, inputs, and technologies are within the abilities of the farmers to own, hire and manage with increasing efficiency. Thus, desirable levels of production in perpetuity with minimal or no adverse effects resources, environment quality and life will be the achievements of this concept. On the other hand, the US Code defined “sustainable agriculture” as an integrated system of plant and animal production practices having a site-specific application that will over the long term: (1) satisfy human food and fiber needs; (2) enhance environmental quality and the natural resource base upon which the agricultural economy depends; (3) make the most efficient use of nonrenewable resources and on-farm resources and integrate, where appropriate, natural biological cycles and controls; (4) sustain the economic viability of farm operations; and (5) enhance the quality of life for farmers and society as a whole (US Senate Committee on Agriculture, Nutrition, & Forestry, 1990). However, there is a general agreement that interactions between farming systems, soil, water, biota, and atmosphere are more complex than they seem, and much more is needed to be understood about their long-term interactions, and often environmental issues are intertwined with economic and social concerns. It is well admitted that our natural resources are limited, and environmental degradation progressed dangerously, hence the necessity in finding a solution to how to feed the growing population.

Indeed, sustainable agriculture depends greatly on natural resources’ preservation, environment quality, and management and maintenance of ecosystems and soils. Because crops and animal products are the unique source of food, sustainability became one of the major contemporary concerns. With the incommensurate development of biological sciences, the outcomes of metabolomics—one among other omics technologies—are further translated to agricultural

research and production technologies, and will help to examine the availability of the natural resources, their sustainability, and their efficient management. This translation has shown its great efficiency in examining how natural resources and environment would lead to reach the goal of sustainable agriculture. Thus, the principles regulating modern agricultural and food production systems include soil management and environment preservation and are considered fundamental rules to the philosophy of sustainability.

During the 20th century, the effects of intensive agricultural production on resource degradation have been of intense concern. Thus, the term of “sustainable agriculture” began to enter the vocabulary of the people responsible for agricultural research resource allocation (Dorfman, 1991; Ruttan, 1992). The present period is known as an important “turn” for agriculture and the trend is to move from an industrial to a sustainable agriculture which focuses on the ability to sustain the growth in the demands and this is considered maybe the most remarkable transition in the history of agriculture (Dorfman, 1991; Horrigan, Lawrence, & Walker, 2002). Because our future depends primordially on the wellbeing of our planet and what it offers us, we need to think of new opportunities for advancing agricultural research to reverse the urgency of the concerns. Since advances in metabolomics are occurring rapidly, they should be imminently translated to agricultural research and production technologies, and this translation could be a key answer to address these concerns.

8.2 Sustainable agriculture and agro-production systems

The agro-production systems offer a valuable entry point in exploring the challenge of achieving sustainable production patterns, and the goals of sustainable agro-systems aim to improve the quality of life while ensuring social, environmental, and economic sustainability. Indeed, the greatest challenge is to protect and sustainably manage the natural resource, and in parallel feeding and housing the growing world population.

Intensive agriculture (IA) tends to produce cheaper and more abundant food crops; however, the environmental impacts and social costs of this agriculture are increasingly apparent.

First, IA has significantly contributed to degrading soils, and the area of cropland per capita has been steadily declining, from 0.43 ha in 1961 to about 0.24 ha in 2004 (UN, 1991; UNDP, 1998). From the 1950s, c. 5–6 million hectares are lost each year due to soil degradation, and c. 2 billion hectares became uncultivated because of poor agricultural practices (WRI, 1998), and, the dramatic rise in grain yields during the 1960s and 1970s tended to outweigh the loss of arable land. On the other hand, rather than protection arable land, millions of hectares of cropland continue to be lost at a serious rate due to urbanization and building new cities.

Second, IA is also associated with extensive chemical inputs and the resulting higher yield outputs are the depletion of soils fertility, creation of pesticide-resistant pests, and environmental and health impacts. For example, the use of nitrogen fertilizers increased by fivefold during the 40 last years; contaminating surface and groundwater with nitrates presently considered the most common chemical contaminants in drinking water. Of course, organic farming (OF) has been seen as a good approach to dramatically reduce chemicals dependency, improve the fertility and structure of soils by using crop rotation, recycling crop residues, and applying organic manures and mulches. OF system has also shown to protect surface and groundwater and help preserve biodiversity, making this agroecological system more conserving of the environment and natural resources. Indeed, this system is based on different pillars consisting of integrating natural processes such as regeneration and minimizing the use of nonrenewable inputs (pesticides and fertilizers).

Unfortunately, only a small percent of agricultural producers are organic farmers, and the challenge is to find the strategy by which sustainable agriculture can successfully make the transition to become the standard operating framework for producing food and achieving sustainable food security. Biotechnology is another controversial development of industrial agriculture by integrating genetically modified (GM) food into the global food system. According to the FAO (2000), the application of biotechnology to agriculture might enhance food security; however, GM organisms are still under intense debate.

8.3 Concepts of metabolomics and their applications to agriculture

The term metabolomics derives from metabolites and is resulting from the interaction of the system’s genome with its environment (GxE), and this discipline entails the study of global metabolite profiles in a system (cell, tissue, or organism) under a given set of conditions (Fell & Wagner, 2000; Glassbrook, Beecher, & Ryals, 2000; Goodacre, Vaidyanathan, Dunn, Harrigan, & Kell, 2004; Mitchell, Holmes, & Carmichael, 2002; Schmidt, 2004). For simplicity, metabolomics aims to “measure the time-related multiparametric metabolic responses of a biological system to a

physiological or environmental stimulus or genetic modification” (Mitchell et al., 2002; Nicholson, Lindon, & Holmes, 1999). By analyzing the complete range of metabolites present within a biological system, a clear biological face of this system at a defined developmental stage or under specific environmental factors should be reflected (Brown et al., 2005). To achieve this, several comprehensive technologies, such as chromatographic techniques coupled to mass-spectrometry are employed to study these wide arrays of metabolites (Dunn & Ellis, 2005). Metabolomics was first described in the 1950s, and afterward, it developed during the four following decades. With the development of analytical technologies, metabolomics knew a tremendous development during the last 20 years, and it became an area of major research interest.

From the ancient time, plants have provided a major part of human food, and cereals and other starchy plants are a major source of energy. To feed and support the growing human population, plant breeding focused on increasing the yield per hectare and crops protection from pests and diseases and also reduces losses during cultivation, harvesting, handling, transportation, and storage. The first research on horticultural crops aimed to improve their utility for processing rather than their nutritional values. However, in the early 1960s with the discovery of new varieties and the development of metabolomics and other omics technologies, these aims were extended to physiological and biochemical attributes (Mertz, Bates, & Nelson, 1964; Rochfort, 2005). These technologies and their products are of crucial importance within the context of sustainable development and agricultural genomics, especially for breeding programs in crop and crop genetic engineering and overcoming biotic and abiotic stresses are one of the greatest ultimate impacts (Fig. 8.1). Therefore sustainable agriculture should be considered as a “key target” in the “gene revolution” since the cost of developing new concepts is much less compared to the cost of “damage reparation” caused by these stresses.

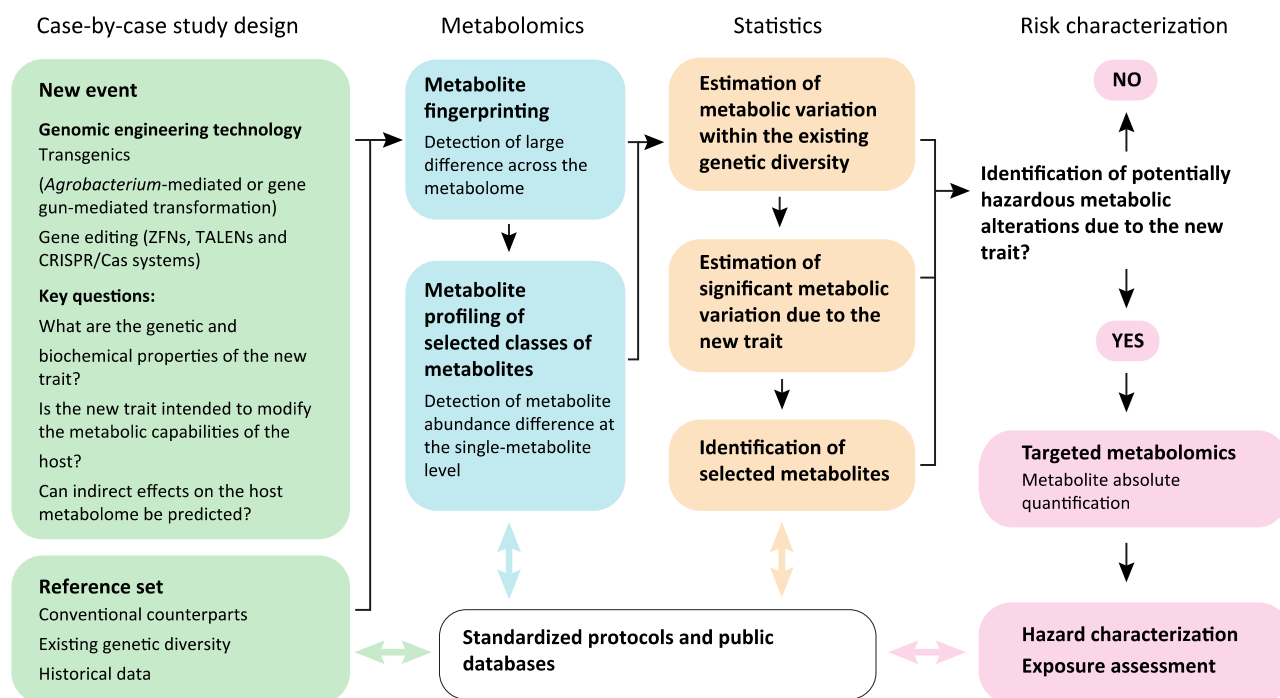


FIGURE 8.1 A proposed framework to integrate untargeted metabolomic analysis in biotech crops risk assessment. This multitiered framework is subdivided into four main stages. An initial collection of the available biological information regarding a new trait allows an assessor to predict potential effects on the crop metabolome (if any). An exhaustive set of references representative of the existing genetic diversity of a crop species is included in the study design to estimate natural metabolic diversity within the species. Information collected in the first stage is used to guide the choice of a dedicated protocol of untargeted metabolomic analysis (metabolite fingerprinting and/or profiling). Statistical analysis of the metabolomic data and partial identification of metabolite features lead to an initial characterization of the potential risk linked to the identified metabolic alterations. This initial risk characterization can then trigger further in-depth hazard characterization and exposure assessment. *CRISPR*, Clustered regularly interspaced palindromic repeats; *TALENs*, transcription activator-like effector nucleases; *ZFNs*, zinc finger nucleases. From Christ, B., Pluskal, T., Aubry, S., & Weng, J. K. (2018). Contribution of untargeted metabolomics for future assessment of biotech crops. *Trends in Plant Science*, 23, 1047–1056, with permission of Elsevier.

8.4 Bridging metabolomics to sustainable agriculture

Comprehensively, metabolomics is considered a valuable approach to understand growth, development, response to stresses, and resistance and resilience of plants to environmental changes. By understanding the multiple mechanisms regulating these numerous processes, more efficient strategies can be developed to improve crops production, soil conservation, and, therefore, ensure the sustainability of the agroecosystems.

8.4.1 Metabolomics for biotic and abiotic stresses assessment

The agronomic fall of sustainable agriculture can be biotic or abiotic (Jahangir, Abdel-Farid, Kim, Choi, & Verpoorte, 2009; Shao, Chu, Abdul Jaleel, & Zhao, 2008; Shao et al., 2007). These stresses cause significant losses in crops and might significantly affect productivity. Thus metabolomics is one of the scientific strategies to elucidate the molecular genetic basis of stress response and resistance of crops (Fig. 8.2). By deciphering the mechanisms regulating the

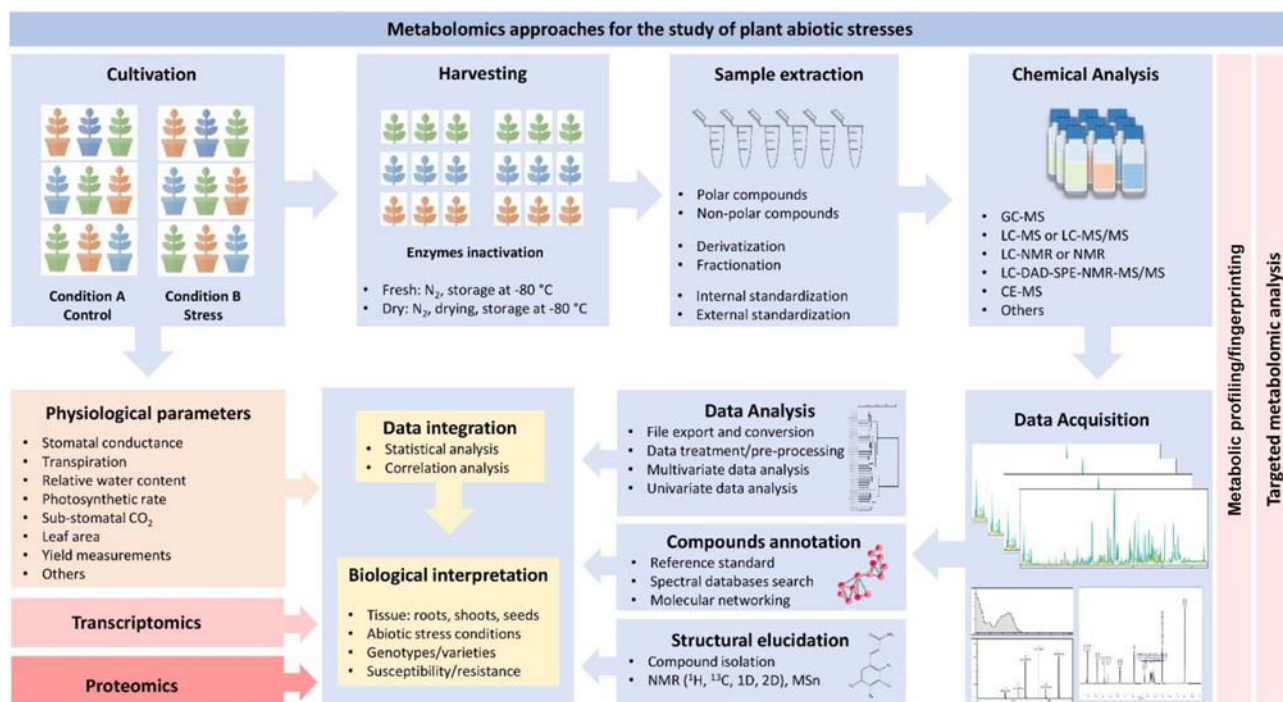


FIGURE 8.2 Schematic representation of the suggested experimental workflow for the metabolomics-assisted study of crops and abiotic stresses. The process starts with the cultivation experiments, which must include at least two different conditions (e.g., stress and control) and a representative number of biological replicates. Depending on the study, different genotypes, varieties, or mutants, susceptible or tolerant, can be arranged and exposed to the experimental conditions. As pointed out by Sanchez et al. (2012), 19 more than two tolerant and sensitive species/ cultivars should be included to avoid a misunderstanding between natural variation and metabolic tolerance. During this phase, the physiological parameters can be monitored and registered. The next step is harvesting. The plant material (shoots, roots, seed, flowers, stems, or others) is harvested and promptly frozen in liquid nitrogen to avoid enzymatic reactions and degradations. In the sequence, the samples can be stored in a freezer at -80°C , dried (usually freeze-dried), or directly extracted from the fresh tissue. Before extraction, the samples must be powdered, homogenized, and weighted. The best extraction protocol must be chosen according to the desired purpose (e.g., considering targeted metabolomics analysis or metabolic profiling/fingerprinting) and also considering the different classes of metabolites that can be extracted. Usually, internal standardization is required for subsequent normalizations and data analysis. Then, samples are subjected to chemical analysis (using different analytical platforms). In general, most of the metabolomics protocols include a separation step (by LC or GC, mainly) hyphenated to the detection technique of choice (usually Ms or NMR in different arrays). After data acquisition, the raw files are exported for data analysis. The high-throughput process considers several steps such as the conversion to suitable formats, preprocessing, normalizations, data cleaning, alignment, and corrections, among others. Multivariate data analysis methods can be used to evaluate the quality of the acquired data. Additionally, compounds can be annotated by comparing the obtained spectra with those available in mass spectral reference libraries. Still, if necessary, the compounds can be identified by complete structural elucidation (which requires, most of the time, isolation and purification). During this process, the information can be analyzed by different statistical, univariate, or multivariate data analysis tools. Finally, the metabolomics results can be integrated with transcriptomics or proteomics data and/or with the corresponding physiological data for biological interpretation. From Bueno, P. C. P., & Lopes, N. P. (2020). *Metabolomics to characterize adaptive and signaling responses in legume crops under abiotic stresses*. ACS Omega, 5, 1752–1763, with permission under an ACS AuthorChoice License.

expression of stress-related genes, scientists would have better insights into fundamental issues in plant biology which are required for the genetic improvement of food crops (Dita, Risipail, Prats, Rubiales, & Singh, 2006).

So far, water limitation is likely the major concern for agriculture and the concern is worsened by climate change since drought is among the effects of impending this change. Water scarcity is predicted to have a profound impact on crop productivity and yield, and will affect many regions worldwide (Shanker et al., 2014). On the other hand, plants developed different strategies to cope with water scarcity, and one of these mechanisms is the production and accumulation of organic compounds which act as osmoregulators, osmoprotectants, and/or turgor maintenance. To improve tolerance and coping of plants with drought, a deeper understanding and elucidation of the metabolic pathways producing and regulating these compounds constitutes one of the most promising alternatives (Bueno & Lopes, 2020). Therefore, metabolomics approach is considered an effective strategy to explore system responses to drought. Interestingly, drought stress was the most investigated stress comparatively to other abiotic stresses and extensive literature is readily available on the subject. For example, metabolomics analysis showed abundant changes in amino acids, organic acids, sugars, and phenolic compounds under water deficiency in wheat (Michaletti, Naghavi, Toorchi, Zolla, & Rinalducci, 2018), sesame (You et al., 2019), eggplant (Mibei, Owino, Ambuko, Giovannoni, & Onyango, 2018), cowpea (Goufo et al., 2017), pea (Charlton et al., 2008), rice (Ma et al., 2016), soybean (Das, Rushton, & Rohila, 2017), oat (Sánchez-Martín et al., 2015), barley (Chmielewska et al., 2016; Hong, Ni, & Zhang, 2020), pepper (Vílchez, Niehaus, Dowling, González-López, & Manzanera, 2018), sugarcane (Budzinski, de Moraes, Cataldi, Franceschini, & Labate, 2019), peanut (Gundaraniya, Ambalam, & Tomar, 2020), maize (Yang et al., 2018), and white clover (Li et al., 2019). Indeed, these and future studies will be imperative in understanding drought to maintain yield and productivity, and therefore, secure food supplies for the future.

Pests and diseases are the major biotic stresses affecting cultivated crops, and their severity and effects are variable, while some of them affect large areas and cause considerable losses in quantity and quality (Nene & Reddy, 1987; Rubiales, Emeran, & Sillero, 2002; Warkentin, Rashid, & Xue, 1996). Soilborne diseases are also very common in crops and most of them attack seedlings causing damping-off and might result in up to 80% of plants death (Denman, Knox-davis, Calitz, & Lamprecht, 1995; Kolkman & Kelly, 2003; Navas-Cortés, Hau, & Jimenez-Diaz, 2000; Wang, Hwang, Chang, Turnbull, & Howard, 2003; Wang, Okamoto, Xing, & Crawford, 2003). On the other hand, viral diseases cause heavy losses for most crops, and many have been considered the most important yield-limiting factor (Coyne et al., 2003), while insects cause important damages both through direct feeding or by transmission of pathogens (Garza, Cardona, & Singh, 1996; Romero-Andreas, Yandell, & Bliss, 1986; Yoshida, Cowgill, & Wightman, 1997). Abiotic stresses such as cold, drought, waterlogging, and salinity affect c.a. 90% of arable lands which experience one or more environmental stresses, and drought, extreme temperature, and high salinity have been shown to dramatically limit crops productivity (Dita et al., 2006). Water deficit constitutes the major abiotic factor affecting crops productivity (Sharma & Lavanya, 2002), while waterlogging causes severe yield losses (Dennis et al., 2000) by limiting O₂ diffusion of the soil and subsequently inducing denitrification and/or nitrate ammonification (Laanbroek, 1990), hence, limiting potassium, sodium, iron, and manganese uptake by plants which become more susceptible to diseases (McDonald & Dean, 1996).

As an efficient approach, metabolomics has been used to assess the involvement of subsets of metabolites in various stresses (Urano, Kurihara, Seki, & Shinozaki, 2010). Several research works focused on phenolics accumulation observed in response to pathogens infection (Baldrige, O'Neill, & Samac, 1998; Borejsza-Wysocki, Borejsza-Wysocka, & Hrazdina, 1997; Lozovaya, Lygin, Li, Hartman, & Widhohn, 2004; Saunders & O'Neill, 2004; Shimada, Akashi, Aoki, & Ayabe, 2000). Many phenolic and other compounds have been described as potential defense or signal molecules such as terpenoids (He & Dixon, 2000; Mithofer, Muller, Wanner, & Eichacker, 2002), for example, copper or mercury stress induced the accumulation of phytoalexin (Wu & VanEtten, 2004). Although significant advances have been achieved, still the roles of these molecules remain unclear and not fully elucidated. However, an interesting hypothesis suggests that phenolic compounds act as attractants for natural predators of herbivorous insects and as systemic signal defense. Nowadays, large-scale metabolomics analysis is providing large data sets helping in identifying potential marker candidates to increase intrinsic resistance and tolerance levels of crops to these stresses.

8.4.2 Metabolomics for soils science and soil conservation

The term “macronutrient” refers to one of the nine elements needed by plants in larger quantities, namely, nitrogen, phosphorus, sulfur, calcium, magnesium, potassium, carbon, hydrogen, and oxygen. However, only the first six that can be supplied conveniently as fertilizers are ordinarily considered among the macronutrients (Broyer & P Stout, 1959). Soil fertility is a continuous process beginning before crops establishment, and aiming to adjust soils conditions for an optimal growth of plants, and this management is the key to sustainable agriculture. Although it seems simple, the concept of soil fertility is difficult to explain, and some nutrients are more accessible in some soils than others, although these later contain higher fertility resources (Cakmak, 2002; Loneragan, 1997; Tisdale, Nelson, & Beaton, 1985).

The availability of nitrogen is considered as the major limiting factor in agricultural productivity, and its assimilation is necessary for plant growth and development (Newbould, 1989). Hence, one of the driving forces behind agricultural sustainability is the effective management of nitrogen through the biological nitrogen fixation (Bohlool, Ladha, Garrity, & George, 1992; Vance & Graham, 1994), and 80% of this biologically fixed nitrogen which become available to crops is performed by the soil bacteria, namely, *Rhizobium*, *Bradyrhizobium*, *Sinorhizobium*, and *Azorhizobium*-Legume symbiosis (Peoples, Herridge, & Ladha, 1995; Vance, 1996, 1997). Phosphorus (P) is the second important mineral for crops, and is present as mineral deposit, and presents the positive point to remain in soils. Phosphorus fixation often requires the addition of excess P fertilizers to meet the requirements of crops (Brady, 1990; Loneragan, 1997); however, only 15%–20% of the applied quantity would become available to crops (Prasad & Power, 1997). In soils, potassium (K) is present in much large amount and fairly well distributed in soils (Prasad & Power, 1997), under four different forms (Brady, 1990; Mulder, 1950; Sparks, 1987; Tisdale et al., 1985). Nevertheless, little is known on potassium but genotypic difference in crop species with respect to potassium nutrition has been reported (Glass & Perley, 1980). Calcium (Ca) was recognized a long time ago necessary for the growth of higher plants (Vance, 1997) and required for structural roles in the cell wall and membranes, as well as its role as a counter-cation for inorganic and organic anions in the vacuole, and as an intracellular messenger in the cytosol (Marschner, 1995; Martinoi, Maeshima, & Neuhaus, 2007). However, the required levels of Ca and the tolerance of plants to their concentration in the rhizosphere differ from species to species (Hepler, 2005; Martinoi et al., 2007; McLaughlin & Wimmer, 1999; Plieth, 2005; White & Broadley, 2001). Similarly to calcium, magnesium (Mg) is fairly mobile in plants and the most abundant divalent cation in a living cell, and share few traits with calcium such as (1) both are taken up by plants as cations, and (2) they are basic or basic forming elements (Tisdale et al., 1985). Magnesium plays an even more prominent role in plants as an essential component in the structure of chlorophyll and ribosomes; however, its uptake, transport, and homeostasis in eukaryotes are poorly understood (Li, Tutone, Drummond, Gardner, & Luan, 2001; Matsumoto, 2000; Prasad & Power, 1997; Rengel & Robinson, 1989; Tan, Keltjens, & Findenegg, 1991). Sulfur (S), another macronutrient, is the least abundant in plants, and required in low levels by plants for their growth since it has numerous biological functions (Leustek, Martin, Bick, & Davies, 2000; Schmidt & Jäger, 1992), for example, the biosynthesis of primary and secondary metabolites and coenzymes (Giovannoni, 2004; Hell, 1997; Leustek et al., 2000; Schmidt & Jäger, 1992).

One of the objectives of metabolomics approaches aims also to integrate with plant nutrition research to develop new plant genotypes with greater uptake and acquisition efficiency of nutrients from soils, hence, contributing greatly to reduce the massive addition of fertilizers. Imbalanced levels of essential or putatively nonessential microelements are potentially toxic, and soils are contaminated as a result of mining or industrial and domestic activities and fertilization (Cheng, 2003; di Toppi & Gabbrielli, 1999; Meharg, 2004; Nriagu & Pacyna, 1988). Some interesting studies have been carried out to understand the effects of some nutrients' uptake by plants during their growth and development. Roessner et al. (2006) used a metabolomics approach to assess the effect of boron (B) on plant growth at either deficient or toxic concentrations in soil by comparing metabolite profiles in root and leaf tissues. Postgenomic investigations have also been carried out to study the metabolism of sulfur (Hirai & Saito, 2004), and the metabolome of leaf and root samples was analyzed (Hirai et al., 2004). Because the pattern of the gene expression is regulated by a metabolite accumulation pattern and vice versa, metabolomics approaches could also be considered indispensable for the understanding of the whole mechanism of sulfur (Hirai & Saito, 2004). The physiological responses of phosphorus homeostasis were also evaluated by the shoot ionome of plants grown under different P conditions, and the results showed that multivariable ionomics signatures are associated with mineral nutrient homeostasis (Baxter et al., 2008). Although barely comparable, cadmium (Cd), one of the impurities of phosphatic fertilizers and one of the contaminants in soils, affects the physiological condition of plants (Koepppe, 1977), and metabolomics contributed to understanding the consequences of exposure of plants to cadmium (Cd) (Bailey, Oven, Holmes, Nicholson, & Zenk, 2003).

In 1904 Lorenz Hiltner was the first agronomist coining the term emphasizing the critical role of rhizosphere microbial activities in the nutrition and health of plants and stating the nutrition of plants depends on the composition of the soil flora in this soil system (Hartmann, Rothballer, Schmid, & Hiltner, 2008). The rhizosphere is the volume of soil influenced by the root and the root tissues colonized by microorganisms which react to the metabolites released by plant roots, and these interactions influence plants growth and development, change nutrient dynamics, and alter plants' susceptibility to diseases and abiotic stresses (Barea, Pozo, Azcón, & Azcón-Aguilar, 2005; Linderman, 1988; Lynch, 1987; Morgan & Whipps, 2001; Morgan, Bending, & White, 2005; Pinton, Varanini, & Nannipieri, 2001; Smith, 2002). The perpetual functioning of the rhizosphere ecosystem is crucial for soil sustainability and productivity, and understanding the processes occurring in this ecosystem is a key element in soils management practices and enable better decision-making for sustainability concept (Schloss & Handelsman, 2003; Van Elsas, Wellington, & Trevors, 1997). Consequently, these interactions have considerable potential for metabolomics exploration, however, identification,

quantification, and measurement of the processes occurring in the rhizosphere are often difficult or tedious (Barea et al., 2005; Johnson, Ijdo, Genney, Anderson, & Alexander, 2005).

8.4.3 Metabolomics for crops production

With the increasing world populations and more people moving from rural to urban areas, food production and food security are becoming a task in reducing hunger, and malnutrition, and the consequences of climate change will make this task even more difficult than it is today particularly in some regions of the world such as Africa and some Asian countries. From the end of the 20th century, much efforts and new approaches have been put into plant breeding to increase crops productivity and yield, and enhance the resistance of crops to pests and diseases that cause heavy damage and losses during cropping, harvesting, handling, transportation, and storage. With the development of new analytical techniques and technologies, these approaches began to change in the early 1960s with the discovery of new varieties (Bailey et al., 2003; Baldridge et al., 1998; Barea et al., 2005; Baxter et al., 2008; Benkeblia, 2012; Bohlool et al., 1992; Borejsza-Wysocki et al., 1997; Brady, 1990; Broyer & P Stout, 1959; Budzinski et al., 2019; Bueno & Lopes, 2020; Cakmak, 2002; Charlton et al., 2008; Cheng, 2003; Chmielewska et al., 2016; Christ, Pluskal, Aubry, & Weng, 2018; Coyne et al., 2003; Das et al., 2017; Denman et al., 1995; Dennis et al., 2000; di Toppi & Gabbriellini, 1999; Dita et al., 2006; Garza et al., 1996; Giovannoni, 2004; Glass & Perley, 1980; Goufo et al., 2017; Gundaraniya et al., 2020; Hartmann et al., 2008; He & Dixon, 2000; Hell, 1997; Hepler, 2005; Hirai & Saito, 2004; Hirai et al., 2004; Hong et al., 2020; Jahangir et al., 2009; Johnson et al., 2005; Koeppe, 1977; Kolkman & Kelly, 2003; Laanbroek, 1990; Leustek et al., 2000; Li et al., 2001, 2019; Linderman, 1988; Loneragan, 1997; Lozovaya et al., 2004; Lynch, 1987; Ma et al., 2016; Marschner, 1995; Martinoi et al., 2007; Matsumoto, 2000; McDonald & Dean, 1996; McLaughlin & Wimmer, 1999; Meharg, 2004; Mibei et al., 2018; Michaletti et al., 2018; Mithofer et al., 2002; Morgan & Whipps, 2001; Morgan et al., 2005; Mulder, 1950; Navas-Cortés et al., 2000; Nene & Reddy, 1987; Newbould, 1989; Nriagu & Pacyna, 1988; Peoples et al., 1995; Pinton et al., 2001; Plieth, 2005; Prasad & Power, 1997; Rengel & Robinson, 1989; Rochfort, 2005; Roessner et al., 2006; Romero-Andreas et al., 1986; Rubiales et al., 2002; Sánchez-Martín et al., 2015; Saunders & O'Neill, 2004; Schloss & Handelsman, 2003; Schmidt & Jäger, 1992; Shanker et al., 2014; Shao et al., 2007, 2008; Sharma & Lavanya, 2002; Shimada et al., 2000; Smith, 2002; Sparks, 1987; Tan et al., 1991; Tisdale et al., 1985; Urano et al., 2010; Vance & Graham, 1994; Vance, 1996, 1997; Van Elsas et al., 1997; Vílchez et al., 2018; Wang, Hwang et al., 2003; Wang, Okamoto et al., 2003; Warkentin et al., 1996; White & Broadley, 2001; Wu & VanEtten, 2004; Yang et al., 2018; Yoshida et al., 1997; You et al., 2019). Indeed, the modern crop science production aims to broaden our still limited knowledge on one hand, how genes can be “shaped” and consequently affect enzymes and metabolites, and on the other hand, gene–enzyme–metabolite interactions and how their effects lead to “new” crops possessing desirable agronomic, physiological, biochemical, and nutritional features. Indeed, the “road is still long ahead” and the problem huge; however, metabolomics, and more generally omics technologies, are giving us hope and researchers have started piecing together interesting clues (Alawiye and Babalola, 2021). These modern technologies led to give new insights into the mysteries of genes expression and function, metabolism and metabolic pathways, and, more broadly, genetic diversity within and between plants, their responses to biotic and abiotic stresses, and their limits in genetic modifications (Benkeblia, 2011). The achievements of these technologies have resulted in the generation of new research areas devoted to “looking from inside the smaller” rather than looking from the outside (Powell, 2007; Rezzi, Ramadan, Fay, & Kochhar, 2007; Rezzi, Ramadan, Martin et al., 2007; Rist, Wenzel, & Daniel, 2006; Subbiah, 2006; Trujillo, Davis, & Milner, 2006). On the other hand, consequently to these new emerging methodologies, new possibilities are still arising to account for tailored food production methods to increase yield and productivity of crops under adverse conditions. Therefore the good examples of the application of metabolomics are to combine it with molecular biology, chemistry, agriculture, and food science to develop transgenic or GM crops (Chen & Lin, 2013; Engel, Frenzel, & Miller, 2002; Herrera-Estrella, 2000; McGloughlin, 2010; Phillips, 2008; Ruth, 2003; Schilter & Constable, 2002). Besides these goals, metabolomics has also shown a great importance in the analysis of genetic variation and traceability of crops, fruit development, maturation, and ripening and also in elucidating the response and resilience of crops to the environment and stresses (Sousa Silva, Cordeiro, Roessner, & Figueiredo, 2019).

8.4.4 Metabolomics for crops quality

The food qualities and nutritional points of sustainable agriculture falls include content and quality of macronutrients such as starch, protein, oil, and micronutrients (Mazur, Krebbers, & Tingey, 1999) and metabolomics could be considered as one of the foundations of sustainable agriculture and environment care (Campbell, Brunner, Jones, & Strauss,

2003; Fernie & Schauer, 2009; Mazur et al., 1999; Somerville & Somerville, 1999). Metabolomics, combined with other omics technologies, were used to enhance contents of essential and nonessential macronutrients and micronutrients, such as vitamins (e.g., A, C, E, folate), minerals (e.g., iron and zinc), and proteins (Beyer et al., 2002; Oresic, 2009; Potrykus, 2001; Wu, Chen, & Folk, 2003; Zarate, 2010), while vitamin pathways have been designed for the synthesis of many other “nonessential” compounds and macronutrients (DellaPena, 1999, 2001). Microbiology, both food and general, is also profiting from metabolomics to study microbial metabolism which is of special interest concerning human and animal health (Grivet, Delort, & Portais, 2003). Microbial metabolome also includes applications in basic studies of foodborne pathogens and foods metabolism resulting from its microbial ecology such as cheeses ripening (Rager, Binet, Ionescu, & Bouvet, 2000).

More recently, because foods or ingredients derived from GM crops were perceived disparagingly by consumers because of concerns about unintended effects on human health (Frewer et al., 2004), risk assessment of potential adverse effects on humans and the environment became a necessity since these changes are connected to changes in metabolite levels in plants. Thus, metabolomics is offering a good tool for comparing conventional cultivars to those genetically modified (Risher & Oksman-Caldentey, 2006). For example, Catchpole et al. (2005) reported that the application of metabolomics methodology has been shown to be useful for the investigation of compositional similarity in GM potatoes.

8.4.5 Metabolomics and postharvest crops science

Senescence is a natural process affecting fresh crops (Sacher, 1973), and metabolomics has been recently used to investigate this process. To have a good insight into the metabolic mechanisms involved in fruit senescence, Yun et al. (2016) compared the metabolic pattern of litchi pericarp and their findings showed that senescence is mainly resulting from an oxidative process induced by abscisic acid including lipids, polyphenols, and anthocyanins oxidation. On the other hand, genomics studies showed that gene expression resulting from crop–diseases interactions, physiological disorders, biotic and abiotic stresses, or other inducing phenomena during the postharvest life of fresh commodities trigger the formation of hundreds of different metabolites (Ding et al., 2015; Pech, Purgatto, Girardi, Rombaldi, & Latché, 2013). Some of these metabolites are elicited during specific stresses and involved in accelerating undesirable disorders such as chilling injury (CI) and browning and senescence, while other elicited compounds enhance the resistance of commodities to stresses. Thus, metabolomics profiling is a good tool to identify the elicited marker metabolites resulting from different reactions, leading to the development of optimal conditions of handling, minimally processing, and storage of fresh crops by either diverting the metabolism toward desirable pathways or decelerating the production of undesirable metabolites (Fig. 8.3) (Benkeblia, 2012; Hertog et al., 2011). By integrating metabolomics and storage technologies, this knowledge will be facilitated, and further information will be available on the responses of fresh crops during the postharvest life from harvesting to the table (Fig. 8.4). Climacteric fruits are the most perishable commodities and ripen after harvesting and this physiological process triggered by ethylene burst leads to the modification of color, firmness, texture, aroma, and nutritional quality attributes of crops (Barry & Giovannoni, 2007; Bleecker & Kende, 2000; Giovannoni, 2004). Postharvest metabolic changes in fresh crops have been extensively reported, but limited literature is available on metabolite profiles. Interestingly, mannose, citrate, malate, gluconate, and keto-L-gluconate were identified as ripening marker metabolites of tomato (Oms-Oliub et al., 2011). In avocado, the profiled metabolome showed a correlation between amino acids and differential accumulation of linoleic acid during its ripening (Pedreschi et al., 2014). Similarly, among the 46 identified metabolites in sapodilla (*Manilkara zapota*), 20 showed significant differences during the ripening process (Das & De, 2015), and in capsicum (*Capsicum annum*) carbohydrates pool and their derivatives significantly changed during the ripening (Aizat et al., 2014). In another study, untargeted metabolome profile of kiwi fruit showed an increase of mono-, di-, and tri-saccharides during ripening and a decrease in organic acids with loss of neutral pectic side chains (Mack et al., 2017). Metabolomics studies also targeted specific metabolites as ripening markers and showed the role of malate in the metabolism of starch during ripening and softening of tomato fruit (Centeno et al., 2011), and an increase of succinate, γ -aminobutyric acid, and glutamine and a decrease of 2-oxoglutarate in stored citrus (Sun et al., 2013).

Harvested fresh crops are subject to numerous disorders, and biochemical studies have shown their limit in understanding the processes inducing these undesirable changes which deteriorate the organoleptic and visual quality attributes of commodities. So far, CI is the most important physiological disorder affecting particularly tropical fruits. Arabinose, fructose-6-phosphate, valine, and shikimic acid levels were associated with CI tolerance of tomato (Luengwilai, Saltveit, & Beckles, 2012), and the profiled organic volatiles were also used as potential markers of CI in fresh crops. In another study, a positive correlation was also noted between volatiles emission and CI of basil leaves during storage (Cozzolino et al., 2016).

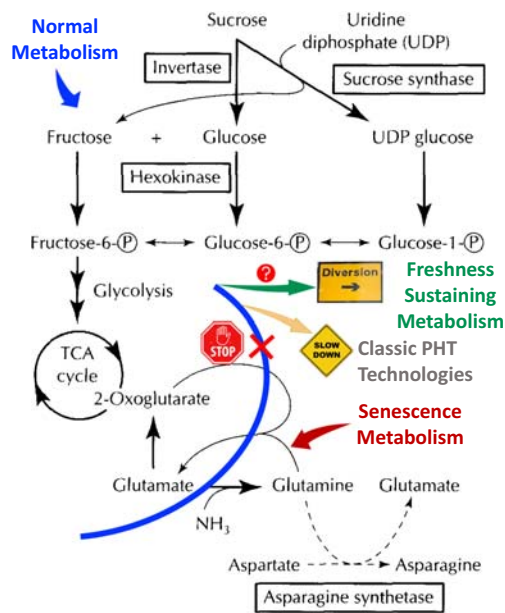


FIGURE 8.3 The senescence metabolism pathway during the postharvest life and the potential roles of metabolomics in extending the shelf-life of fresh crops.

Metabolomics Approach

↓
Understanding the System Biology

↓
Deciphering the Senescence Process

↓
Using Storage Technologies (MAP + Cooling)

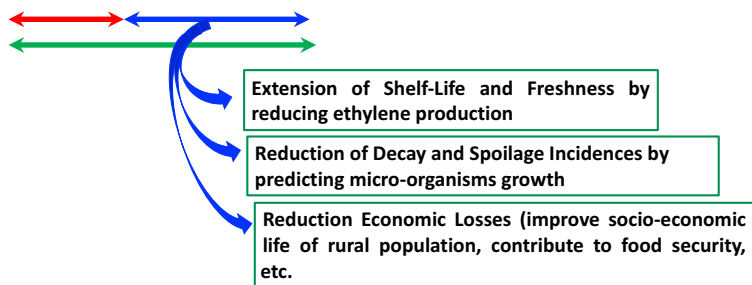


FIGURE 8.4 The role of metabolomics in the development of postharvest technologies to extend the shelf-life of fresh crops.

Other disorders such as superficial scald, braeburn, and mechanical damages might affect commodities during their postharvest life. Superficial scald is a disorder affecting stored pome fruits (Jackman, Yada, Marangoni, Parkin, & Stanley, 1988) and is often associated with CI (Hariyadi & Parkin, 1991; Imahori, Takemura, & Bai, 2008; Sala, 1998) and oxidative stresses. An increase of galactose, mannitol, sorbitol, xylose, and alanine and a decrease of malate and sucrose were noted in apples affected by superficial scald (Hatoum, Annaratone, Hertog, Geeraerd, & Nicolai, 2014). The symptoms of this disorder in apples were also predicted using metabolomics profiling of volatiles, and 6-methyl-5-hepten-2-one, and 6-methyl-5-hepten-2-ol products known to be associated with superficial scald and resulting from α -farnesene, were identified (Rudell, Mattheis, & Hertog, 2009). A similar approach was used to assess the progression of this disorder and three specific volatiles particularly 6-methyl-5-hepten-2-one (MHO) were identified and associated with the symptomatic development of this disorder (Farneti et al., 2015). Furthermore, primary metabolites were profiled after a mechanical impact and results showed noticeable changes in tricarboxylic acid cycle intermediates in potato tuber tissue (Strehmel et al., 2010), and linoleic acid and pentadecanoic acid in mushrooms (O’Gorman, Barry-Ryan, & Frias, 2012), while in cassava higher levels of phenolic acids, scopoletin, carotenoids, and proteins have been identified in physiologically deteriorated tubers (Uarrotta & Maraschin, 2015).

Fresh commodities are also very susceptible to diseases during their handling and storage due to their high moisture and nutrients. Decay and spoilage of fresh crops are observed when symptoms are visible; however, metabolomics approach has been used in an attempt to predict earlier pathogens growth and spoilage (Pinu, 2016), and this approach is based either on host- and pathogen-specific metabolites or the abundance of metabolites (Kushalappa, Vikram, & Raghavan, 2008). In an interesting work, Li, Schmidt, and Gitaitis (2011) profiled volatile organic compounds and interestingly 16 volatiles related to the postharvest spoilage of onion bulb by *Botrytis allii* and *Burkholderia cepacia* were identified as biomarkers of spoilage.

8.5 Conclusions and future perspectives

At the turn of the 21st century, incommensurate progress is being made in the field of metabolomics, and more generally omics technologies, to increase crops production and their resilience to adverse conditions particularly under the changing climate. However, work to date represents a few pieces of the big puzzle in the estimation of an enormous number of species, genomes, and the complex and changing metabolic pathways. Despite the early struggle to understand and elucidate the complex metabolome, metabolomics seems to suit well to tackle many questions. Using metabolomics, many questions are nowadays being answered regarding the metabolome and the functional capabilities of crops and soils, and the alterations in the metabolic profile in response to abiotic and biotic stresses. Because metabolomics become integrative sciences that simultaneously adopt the tools of diverse scientific disciplines and impact diverse scientific areas, the generated data will likely accelerate the discovery of the amazing diversity of the micro-life and give new insights into crops functions, fill the gaps between the discoveries on the “microworld” which are much less compared to those on the “macroworld” because the two are indissociable and each one depends on the other.

Of course, metabolomics is facing many challenges which are diverse and complex. The first challenge is the identification of the crops’ response to abiotic and biotic stresses and identifies the molecular markers of these stresses especially in the field of drought and diseases because this identification will be of great help in developing more resistant and/or resilient varieties. The second challenge is the identification and quantification of the huge number of metabolites of crops, and plants in general. Knowing that some metabolites are found at very low concentrations or have a very short biological life, the analytical techniques should be improved to identify a wider range of metabolites and quantify lower concentrations. This can be achieved by improving the sensitivity and the detection capacity of the analytical techniques which are continuously developed and improved to solve these problems.

Finally, sustainable agriculture does not aim to create new agroecosystems, but it targets preserving the environment and its different systems including the biodiversity, lands, and water. To be simplistic, it is necessary to meet the needs of the present without compromising the needs of the future. Because we cannot remake the world and the pressure on our crop production systems is very high, the wisdom and science recommend to urgently embrace sustainable agriculture to avoid at least unraveling our world.

References

- Aizat, W. M., Dias, D. A., Stangoulis, J. C. R., Able, J. A., Roessner, U., & Able, A. J. (2014). Metabolomics of capsicum ripening reveals modification of the ethylene related-pathway and carbon metabolism. *Postharvest Biology and Technology*, *89*, 19–31.
- Alawiye, T. T., & Babalola, O. O. (2021). Metabolomics: Current application and prospects in crop production. *Biologia (Lahore, Pakistan)*, *76*, 227–239.
- Bailey, N. J. C., Oven, M., Holmes, E., Nicholson, J. K., & Zenk, M. H. (2003). Metabolomic analysis of the consequences of cadmium exposure in *Silene cucubalus* cell culture via 1H NMR spectroscopy and chemometrics. *Phytochemistry*, *62*, 851–858.
- Baldrige, G. D., O’Neill, N. R., & Samac, D. A. (1998). Alfalfa (*Medicago sativa* L.) resistance to the root-lesion nematode, *Pratylenchus penetrans*: Defense-response gene mRNA and isoflavonoid phytoalexin levels in roots. *Plant Molecular Biology*, *38*, 999–1010.
- Barea, J. M., Pozo, M. J., Azcón, R., & Azcón-Aguilar, C. (2005). Microbial cooperation in the rhizosphere. *Journal of Experimental Botany*, *56*, 1761–1778.
- Barry, C. S., & Giovannoni, J. J. (2007). *Ethylene and fruit ripening*, . *Journal of Plant Growth Regulation* (26, pp. 143–159).
- Baxter, I. R., Vitek, O., Lahner, B., Muthukumar, B., Borghi, M., Morrissey, J., et al. (2008). The leaf ionome as a multivariable system to detect a plant’s physiological status. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 12081–12086.
- Benkeblia, N. (2011). *Sustainable agriculture and new biotechnologies*. Boca Raton, FL: CRC Press.
- Benkeblia, N. (2012). Metabolomics and food science: Concepts and serviceability in plant foods and nutrition. In N. Benkeblia (Ed.), *Omics technologies: Tools for food science* (pp. 59–75). Boca Raton, FL: CRC Press.
- Beyer, J. A., Salim, A., Xudong, Y., Lucca, P., Schaub, P., Welsch, R., et al. (2002). “Golden rice”: Introducing the β -carotene biosynthetic pathway into rice endosperm by genetic engineering to defeat vitamin A-deficiency. *The Journal of Nutrition*, *132*, S506–S510.

- Bleecker, A. B., & Kende, H. (2000). Ethylene: A gaseous signal molecule in plants. *Annual Review of Cell and Developmental Biology*, *16*, 1–18.
- Bohlool, B. B., Ladha, J. K., Garrity, D. P., & George, T. (1992). Biological nitrogen fixation for sustainable agriculture: A perspective. *Plant and Soil*, *141*, 1–11.
- Borejsza-Wysocki, W., Borejsza-Wysocka, E., & Hrazdina, G. (1997). Pisatin metabolism in pea (*Pisum sativum* L) cell suspension cultures. *Plant Cell Reports*, *16*, 304–309.
- Brady, N. C. (1990). *The nature and properties of soils* (10th ed.). New York: John Wiley & Sons.
- Brown, M., Dunn, W. B., Ellis, D. I., Goodacre, R., Handl, J., Knowles, J. D., et al. (2005). A metabolome pipeline: From concept to data to knowledge. *Metabolomics: Official Journal of the Metabolomic Society*, *1*, 39–41.
- Broyer, T. C., & P Stout, R. (1959). The macronutrient elements. *Annual Review of Plant Physiology*, *10*, 277–300.
- Budzinski, I. G. F., de Moraes, F. E., Cataldi, T. R., Franceschini, L. M., & Labate, C. A. (2019). Network analyses and data integration of proteomics and metabolomics from leaves of two contrasting varieties of sugarcane in response to drought. *Frontiers in Plant Science*, *10*, 1524. Available from <https://doi.org/10.3389/fpls.2019.01524>.
- Bueno, P. C. P., & Lopes, N. P. (2020). Metabolomics to characterize adaptive and signaling responses in legume crops under abiotic stresses. *ACS Omega*, *5*, 1752–1763.
- Cakmak, I. (2002). Plant nutrition research: Priorities to meet human needs for food in sustainable ways. *Plant and Soil*, *247*, 3–24.
- Campbell, M. M., Brunner, A. M., Jones, M., & Strauss, S. H. (2003). Forestry's fertile crescent: The application of biotechnology to forest trees. *Plant Biotechnology Journal*, *1*, 141–154.
- Catchpole, G. S., Beckmann, M., Enot, D. P., Mondhe, M., Zywicki, B., Taylor, J., et al. (2005). Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 14458–14462.
- Centeno, C. D., Osorio, S., Nunes-Nesi, A., Bertolo, A. L. F., Carneiro, R. T., Araujo, A. L., et al. (2011). Malate plays a crucial role in starch metabolism, ripening, and soluble solid content of tomato fruit and affects postharvest softening. *The Plant Cell*, *23*, 162–184.
- Charlton, A. J., Donarski, J. A., Harrison, M., Jones, S. A., Godward, J., Oehlschlager, S., et al. (2008). Responses of the pea (*Pisum sativum* L.) leaf metabolome to drought stress assessed by nuclear magnetic resonance spectroscopy. *Metabolomics: Official Journal of the Metabolomic Society*, *4*, 312. Available from <https://doi.org/10.1007/s11306-008-0128-0>.
- Chen, H., & Lin, Y. (2013). Promise and issues of genetically modified crops. *Current Opinion in Plant Biology*, *16*, 255–260.
- Cheng, S. (2003). Heavy metal pollution in China: Origin, pattern and control. *Environmental Science and Pollution Research International*, *10*, 192–198.
- Chmielewska, K., Rodziewicz, P., Swarczewicz, B., Sawikowska, A., Krajewski, P., Marczak, Ł., et al. (2016). Analysis of drought-induced proteomic and metabolomic changes in barley (*Hordeum vulgare* L.) leaves and roots unravels some aspects of biochemical mechanisms involved in drought tolerance. *Frontiers in Plant Science*, *7*, 1108. Available from <https://doi.org/10.3389/fpls.2016.01108>.
- Christ, B., Pluskal, T., Aubry, S., & Weng, J. K. (2018). Contribution of untargeted metabolomics for future assessment of biotech crops. *Trends in Plant Science*, *23*, 1047–1056.
- Coyne, D. P., Steadman, J. R., Godoy-Lutz, G., Gilbertson, R., Arnaud-Santana, E., Beaveret, J. S., et al. (2003). Contribution of the bean/cowpea CRSP to management of bean diseases. *Field Crop Research*, *82*, 155–168.
- Cozzolino, R., Pace, B., Cefola, M., Martignetti, A., Stocchero, M., Fratianni, F., et al. (2016). Assessment of volatile profile as potential marker of chilling injury of basil leaves during postharvest storage. *Food Chemistry*, *213*, 361–368.
- Das, A., Rushton, P. J., & Rohila, J. S. (2017). Metabolomic profiling of soybeans (*Glycine max* L.) reveals the importance of sugar and nitrogen metabolism under drought and heat stress. *Plants*, *6*(2), 21. Available from <https://doi.org/10.3390/plants6020021>.
- Das, S., & De, B. (2015). Analyzing changes in metabolite profile during postharvest ripening in *Achras sapota* fruits: GC-MS based metabolomics approach. *International Food Research Journal*, *22*, 2288–2293.
- DellaPenna, D. (1999). Nutritional genomics: Manipulating plant micronutrients to improve human health. *Science (New York, N.Y.)*, *285*, 375–379.
- DellaPenna, D. (2001). Plant metabolic engineering. *Plant Physiology*, *125*, 160–163.
- Denman, S., Knoxdavis, P. S., Calitz, F. G., & Lamprecht, S. C. (1995). Pathogenicity of *Pythium irregulare*, *Pythium sylvaticum* and *Pythium ultimum* var. *ultimum* to Lucerne (*Medicago sativa*). *Australasian Plant Pathology*, *24*, 137–143.
- Dennis, E. S., Dolferus, R., Ellis, M., Rahman, M., Wu, Y., Hoerenet, F. U., et al. (2000). Molecular strategies for improving waterlogging tolerance in plants. *The Journal of Experimental Biology*, *51*, 89–97.
- di Toppi, L. S., & Gabrielli, R. (1999). Response to cadmium in higher plants. *Environmental and Experimental Botany*, *41*, 105–130.
- Ding, Y., Chang, J., Ma, Q., Chen, L., Liu, S., Jin, S., et al. (2015). Network analysis of postharvest senescence process in citrus fruits revealed by transcriptomic and metabolomic profiling. *Plant Physiology*, *168*, 357–376.
- Dita, M., Rispaal, N., Prats, E., Rubiales, D., & Singh, K. B. (2006). Biotechnology approaches to overcome biotic and abiotic stress constraints in legume. *Euphytica*, *147*, 1–24.
- Dorfman, R. (1991). Protecting the global environment: An immodest proposal. *World Development*, *19*, 128, 13.
- Dunn, W. B., & Ellis, D. I. (2005). Metabolomics, Current analytical platforms and methodologies. *Trends in Analytical Chemistry*, *24*, 285–294.
- Engel, K. H., Frenzel, T., & Miller, A. (2002). Current and future benefits from the use of GM technology in food production. *Toxicology Letters*, *127*, 329–336.
- FAO. (2000). *Biotechnology for sustainable agriculture. Report of the Commission on Sustainable Development*. Roma: Food and Agriculture Organization.

- Farneti, B., Busatto, N., Khomenko, I., Cappellin, L., Gutierrez, S., Spinelli, F., et al. (2015). Untargeted metabolomics investigation of volatile compounds involved in the development of apple superficial scald by PTR-ToF-MS. *Metabolomics: Official Journal of the Metabolomic Society*, *11*, 341–349.
- Fell, D. A., & Wagner, A. (2000). The small world of metabolism. *Nature Biotechnology*, *8*, 1121–1122.
- Fernie, A., & Schauer, N. (2009). Metabolomics-assisted breeding: A viable option for crop improvement? *Trends in Genetics: TIG*, *25*, 38–48.
- Frewer, L., Lassen, J. L., Kettlitz, B., Scholderer, J., Beekman, V., & Bermal, K. G. (2004). Social aspects of genetically modified foods. *Food and Chemical Toxicology: An International Journal Published for the British Industrial Biological Research Association*, *42*, 1181–1193.
- Garza, R., Cardona, C., & Singh, S. P. (1996). Inheritance of resistance to bean-pod weevil (*Apion godmani* Wagner) in common beans from Mexico. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *92*, 357–362.
- Giovannoni, J. J. (2004). Genetic regulation of fruit development and ripening. *The Plant Cell*, *16*, S170–S180.
- Glass, A. D. M., & Perley, J. E. (1980). Varietal differences in potassium uptake by barley. *Plant Physiology*, *65*, 160–164.
- Glassbrook, N., Beecher, C., & Ryals, J. (2000). Metabolic profiling on the right path. *Nature Biotechnology*, *18*, 1142–1143.
- Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology*, *22*(5), 245–252.
- Goufo, P., Moutinho-Pereira, J. M., Jorge, T. F., Correia, C. M., Oliveira, M. R., Rosa, E. A. S., et al. (2017). Cowpea (*Vigna unguiculata* L. Walp.) metabolomics: Osmoprotection as a physiological strategy for drought stress resistance and improved yield. *Frontiers in Plant Science*, *8*, 586. Available from <https://doi.org/10.3389/fpls.2017.00586>.
- Grivet, J. P., Delort, A. M., & Portais, J. C. (2003). NMR and microbiology: From physiology to metabolomics. *Biochimie*, *85*, 823–840.
- Gundaraniya, S. A., Ambalam, P. S., & Tomar, R. S. (2020). Metabolomic profiling of drought-tolerant and susceptible peanut (*Arachis hypogaea* L.) genotypes in response to drought stress. *ACS Omega*, *5*, 31209–31219.
- Hariyadi, P., & Parkin, K. L. (1991). Chilling-induced oxidative stress in cucumber fruits. *Postharvest Biology and Technology*, *1*, 33–45.
- Hartmann, A., Rothballer, M., Schmid, M., & Hiltner, L. (2008). A pioneer in rhizosphere microbial ecology and soil bacteriology research. *Plant and Soil*, *312*, 7–14.
- Hatoum, D., Annaratone, C., Hertog, M. L. A. T. M., Geeraerd, A. H., & Nicolai, B. M. (2014). Targeted metabolomics study of 'Braeburn' apples during long-term storage. *Postharvest Biology and Technology*, *96*, 33–41.
- He, X. Z., & Dixon, R. A. (2000). Genetic manipulation of isoflavone 7-O-methyltransferase enhances biosynthesis of 4'-O-methylated isoflavonoid phytoalexins and disease resistance in alfalfa. *The Plant Cell*, *12*, 1689–1702.
- Hell, R. (1997). Molecular physiology of plant sulfur metabolism. *Planta*, *202*, 138–148.
- Hepler, P. K. (2005). Calcium: A central regulator of plant growth and development. *The Plant Cell*, *17*, 2142–2155.
- Herrera-Estrella, L. R. (2000). Genetically modified crops and developing countries. *Plant Physiology*, *124*, 923–926.
- Hertog, M. L. A. T. M., Rudell, D. R., Pedreschi, R., Schaffer, R. J., Geeraerd, A. H., Nicolai, B. M., et al. (2011). Where systems biology meets post-harvest. *Postharvest Biology and Technology*, *62*, 223–237.
- Hirai, M. Y., & Saito, K. (2004). Post-genomics approaches for the elucidation of plant adaptive mechanisms to sulphur deficiency. *Journal of Experimental Botany*, *55*, 1871–1879.
- Hirai, M. Y., Yano, M., Goodenowe, D. B., Kanaya, S., Kimura, T., Awazuhara, M., et al. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 10205–10210.
- Hong, Y., Ni, S. J., & Zhang, G. P. (2020). Transcriptome and metabolome analysis reveals regulatory networks and key genes controlling barley malting quality in responses to drought stress. *Plant Physiology and Biochemistry*, *152*, 1–11.
- Horrigan, L., Lawrence, R. S., & Walker, P. (2002). How sustainable agriculture can address the environmental and human health harms of industrial agriculture. *Environmental Health Perspectives*, *110*, 445–456.
- Imahori, Y., Takemura, M., & Bai, J. (2008). Chilling-induced oxidative stress and antioxidant responses in mume (*Prunus mume*) fruit during low temperature storage. *Postharvest Biology and Technology*, *49*, 54–60.
- Jackman, R. L., Yada, R. Y., Marangoni, A., Parkin, K. L., & Stanley, D. W. (1988). Chilling injury. A review of quality aspects. *Journal of Food Quality*, *11*, 253–278.
- Jahangir, M., Abdel-Farid, I. B., Kim, K., Choi, Y. H., & Verpoorte, R. (2009). Healthy and unhealthy plants: The effect of stress on the metabolism of Brassicaceae. *Environmental and Experimental Botany*, *67*, 23–33.
- Johnson, D., Ijdo, M., Genney, D. R., Anderson, I. C., & Alexander, I. J. (2005). How do plants regulate the function, community structure and diversity of mycorrhizal fungi? *Journal of Experimental Botany*, *56*, 1751–1760.
- Koeppel, D. E. (1977). The uptake, distribution, and effect of cadmium and lead in plants. *Science of the Total Environment*, *7*, 197–206.
- Kolkman, J. M., & Kelly, J. D. (2003). QTL, conferring resistance and avoidance to white mold in common beans. *Crop Science*, *43*, 539–548.
- Kushalappa, K. C., Vikram, A., & Raghavan, G. S. V. (2008). Metabolomics of headspace gas for diagnosing diseases of fruits and vegetables after harvest. *Stewart Postharvest Review*, *2*, 10. Available from <https://doi.org/10.2212/spr.2008.2.10>.
- Laanbroek, H. (1990). Bacterial cycling of minerals that affect plant-growth in waterlogged soils: A review. *Aquatic Botany*, *38*, 109–125.
- Leustek, T., Martin, M. N., Bick, J. A., & Davies, J. P. (2000). Pathways and regulation of sulfur metabolism revealed through molecular and genetic. *Annual Review of Plant Physiology and Plant Molecular Biology*, *51*, 141–165.
- Li, C., Schmidt, N. E., & Gitaitis, R. (2011). Detection of onion postharvest diseases by analyses of headspace volatiles using a gas sensor array and GC-MS. *LWT—Food Science and Technology*, *44*, 1019–1025.

- Li, L., Tutone, A. F., Drummond, R. S. M., Gardner, R. C., & Luan, S. (2001). A Novel family of magnesium transport genes in Arabidopsis. *The Plant Cell*, *13*, 2761–2775.
- Li, Z., Cheng, B., Yong, B., Liu, T., Peng, Y., Zhang, X., et al. (2019). Metabolomics and physiological analyses reveal β -sitosterol as an important plant growth regulator inducing tolerance to water stress in white clover. *Planta*, *250*, 2033–2046.
- Linderman, R. G. (1988). Mycorrhizal interactions with the rhizosphere microflora – The mycorrhizosphere effect. *Phytopathology*, *78*, 366–371.
- Loneragan, J. F. (1997). Plant nutrition in the 20th and perspectives for the 21st century. *Plant and Soil*, *196*, 163–174.
- Lozovaya, V. V., Lygin, A. V., Li, S., Hartman, G. L., & Widhohn, J. M. (2004). Biochemical response of soybean roots to *Fusarium solani* f. sp. *glycines* infection. *Crop Science*, *44*, 819–826.
- Luengwilai, K., Saltveit, M., & Beckles, D. M. (2012). Metabolite content of harvested Micro-Tom tomato (*Solanum lycopersicum* L.) fruit is altered by chilling and protective heat-shock treatments as shown by GC–MS metabolic profiling. *Postharvest Biology and Technology*, *63*, 116–122.
- Lynch, J. M. (1987). *The rhizosphere*. Chichester: Wiley Inter-science.
- Ma, X., Xia, H., Liu, Y., Wei, H., Zheng, X., Song, C., et al. (2016). Transcriptomic and metabolomic studies disclose key metabolism pathways contributing to well-maintained photosynthesis under the drought and the consequent drought-tolerance in rice. *Frontiers in Plant Science*, *7*, 1886. Available from <https://doi.org/10.3389/fpls.2016.01886>.
- Mack, C., Wefers, D., Schuster, P., Weinert, C. H., Egert, B., Bliedung, S., et al. (2017). Untargeted multi-platform analysis of the metabolome and the non-starch polysaccharides of kiwifruit during postharvest ripening. *Postharvest Biology and Technology*, *125*, 65–76.
- Marschner, H. (1995). *Mineral nutrition of higher plants* (2nd ed.). London: Academic Press.
- Martinoi, E., Maeshima, M., & Neuhaus, H. E. (2007). Vacuolar transporters and their essential role in plant metabolism. *Journal of Experimental Botany*, *58*, 83–102.
- Matsumoto, H. (2000). Cell biology of aluminum toxicity and tolerance in higher plants. *International Review of Cytology*, *200*, 1–46.
- Mazur, B., Krebbers, E., & Tingey, S. (1999). Gene discovery and product development for grain quality traits. *Science (New York, N.Y.)*, *285*, 372–375.
- McDonald, G. K., & Dean, G. (1996). Effect of waterlogging on the severity of disease caused by *Mycosphaerella pinodes* in peas (*Pisum sativum* L.). *Australian Journal of Experimental Agriculture*, *36*, 219–222.
- McGloughlin, M. N. (2010). Modifying agricultural crops for improved nutrition. *New Biotechnology*, *27*, 494–504.
- McLaughlin, S. B., & Wimmer, R. (1999). Calcium physiology and terrestrial ecosystem processes. *The New Phytologist*, *142*, 373–387.
- Meharg, A. A. (2004). Arsenic in rice – Understanding a new disaster for South-East Asia. *Trends in Plant Science*, *9*, 415–417.
- Mertz, E. T., Bates, L. S., & Nelson, O. E. (1964). Mutant maize that changes the protein composition and increases the lysine content of maize endosperm. *Science (New York, N.Y.)*, *145*, 279–280.
- Mibe, E. K., Owino, W. O., Ambuko, J., Giovannoni, J. J., & Onyango, A. N. (2018). Metabolomic analyses to evaluate the effect of drought stress on selected African Eggplant accessions. *Journal of the Science of Food and Agriculture*, *98*, 205–216.
- Michaletti, A., Naghavi, M. R., Toorchi, M., Zolla, L., & Rinalducci, S. (2018). Metabolomics and proteomics reveal drought-stress responses of leaf tissues from spring-wheat. *Scientific Reports*, *8*, 5710. Available from <https://doi.org/10.1038/s41598-018-24012-y>.
- Mitchell, S., Holmes, E., & Carmichael, P. (2002). Metabonomic and medicine: The biochemical oracle. *Biologist (London, England)*, *49*, 217–221.
- Mithofer, A., Muller, B., Wanner, G., & Eichacker, L. A. (2002). Identification of defence-related cell wall proteins in *Phytophthora sojae*-infected soybean roots by ESI-MS/MS. *Molecular Plant Pathology*, *3*, 163–166.
- Morgan, J. A. W., Bending, G. D., & White, P. J. (2005). Biological costs and benefits to plant –microbe interactions in the rhizosphere. *Journal of Experimental Botany*, *56*, 1729–1739.
- Morgan, J. A. W., & Whipps, J. M. (2001). Methodological approaches to the study of rhizosphere carbon flow and microbial population dynamics. In R. Pinton, Z. Varanini, & P. Nannipieri (Eds.), *The rhizosphere: Biochemistry and organic substances at the soil–plant interface* (pp. 373–410). New York: Marcel Dekker Inc.
- Mulder, E. G. (1950). Mineral nutrition of plants. *Annual Review of Plant Physiology*, *1*, 1–24.
- Navas-Cortés, J. A., Hau, B., & Jimenez-Diaz, R. M. (2000). Yield loss in chickpea in relation to development of *Fusarium* wilt epidemics. *Phytopathology*, *90*, 1269–1278.
- Nene, Y. L., & Reddy, M. V. (1987). Chickpea diseases and their control. In M. C. Saxena, & K. B. Singh (Eds.), *The chickpea* (pp. 233–270). Oxon: CAB International.
- Newbould, P. (1989). The use of nitrogen fertiliser in agriculture. Where do we go practically and ecologically? *Plant and Soil*, *115*, 297–311.
- Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). Metabonomics: Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica: The Fate of Foreign Compounds in Biological Systems*, *29*, 1181–1189.
- Nriagu, J. O., & Pacyna, J. M. (1988). Quantitative assessment of worldwide contamination of air water and soils by trace metals. *Nature*, *333*, 134–139.
- O’Gorman, A., Barry-Ryan, C., & Frias, J. M. (2012). Evaluation and identification of markers of damage in mushrooms (*Agaricus bisporus*) postharvest using a GC/MS metabolic profiling approach. *Metabolomics: Official Journal of the Metabolomic Society*, *8*, 120–132.
- Okigbo, B. N. (1991). *Development of sustainable agricultural production systems in Africa*. Ibadan: International Institute of Tropical Agriculture.
- Oms-Oliub, G., Hertog, M. L. A. T. M., Van de Poel, B., Ampofo-Asiama, J., Geeraerd, A. H., & Nicolai, B. M. (2011). *Metabolic characterization of tomato fruit during preharvest development, ripening, and postharvest shelf-life*. *Postharvest Biology and Technology* (62), pp. 7–16.

- Oresic, M. (2009). Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. *Nutrition, Metabolism & Cardiovascular Diseases*, 19, 816–824.
- Pech, J. C., Purgatto, E., Girardi, C. L., Rombaldi, C. V., & Latché, A. (2013). Current challenges in postharvest biology of fruit ripening. *Current Agricultural Science and Technology*, 19, 1–18.
- Pedreschi, R., Munoz, P., Robledo, P., Becerra, C., Defilippi, B. G., Eekelen, H. D. L. M., et al. (2014). Metabolomics analysis of postharvest ripening heterogeneity of ‘Hass’ avocados. *Postharvest Biology and Technology*, 92, 172–179.
- Peoples, M. B., Herridge, D. E., & Ladha, J. K. (1995). Biological nitrogen fixation: An efficient source of nitrogen for sustainable agricultural production? *Plant and Soil*, 174, 3–28.
- Phillips, T. (2008). Genetically modified organisms (GMOs): Transgenic crops and recombinant DNA technology. *Nature Education*, 1, 213.
- Pinton, R., Varanini, Z., & Nannipieri, P. (2001). *The rhizosphere: Biochemistry and organic substances at the soil-plant interface*. New York: Marcel Dekker.
- Pinu, F. R. (2016). Early detection of food pathogens and food spoilage microorganisms: Application of metabolomics. *Trends in Food Science and Technology*, 54, 213–215.
- Pliech, C. (2005). Calcium: Just another regulator in the machinery of life? *Annals of Botany*, 96, 1–8.
- Potrykus, I. (2001). Golden Rice and beyond. *Plant Physiology*, 125, 1157–1161.
- Powell, K. (2007). Functional foods from biotech—An unappetizing prospect? *Nature*, 25, 525–531.
- Prasad, R., & Power, J. F. (1997). *Soil fertility management for sustainable agriculture*. Boca Raton, FL: CRC Press.
- Rager, M. H., Binet, M. R. B., Ionescu, G., & Bouvet, O. M. M. (2000). ³¹P and ¹³C NMR studies of mannose metabolism in *Plesiomonas shigelloides*. *European Journal of Biochemistry/FEBS*, 267, 5136–5141.
- Rengel, Z., & Robinson, D. L. (1989). Competitive aluminum ion inhibition of net magnesium ion uptake by intact *Lolium multiflorum* roots. *Plant Physiology*, 91, 1407–1413.
- Rezzi, S., Ramadan, Z., Fay, L. B., & Kochhar, S. (2007). Nutritional metabolomics: Applications and perspectives. *Journal of Proteome Research*, 6, 513–525.
- Rezzi, S., Ramadan, Z., Martin, F. P., Fay, L. B., van Bladeren, P., Lindon, J. C., et al. (2007). Human metabolic phenotypes link directly to specific dietary preferences in healthy individuals. *Journal of Proteome Research*, 6, 4469–4477.
- Risher, H., & Oksman-Caldentey, K. M. (2006). Unintended effects in genetically modified crops: Revealed by metabolomics? *Trends in Biotechnology*, 24, 102–104.
- Rist, M. J., Wenzel, U., & Daniel, H. (2006). Nutrition and food science go genomic. *Trends in Biotechnology*, 24, 1–7.
- Rochfort, S. (2005). Metabolomics reviewed: A new “omics” platform technology for systems biology and implications for natural products research. *Journal of Natural Products*, 68, 1813–1820.
- Rodale, R. (1988). Agricultural systems: The importance of sustainability. *National Forum*, 68, 2–6.
- Roessner, U., Patterson, J. H., Forbes, M. G., Fincher, G. B., Langridge, P., & Bacic, A. (2006). An investigation of boron toxicity in barley using metabolomics. *Plant Physiology*, 142, 1087–1101.
- Romero-Andreas, J., Yandell, B. S., & Bliss, F. A. L. (1986). Inheritance of a novel seed protein of *Phaseolus vulgaris* L. and its effect on seed composition. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 72, 123–128.
- Rubiales, D., Emeran, A. A., & Sillero, J. C. (2002). Rusts on legumes in Europe and North Africa. *Grain Legumes*, 37, 8–9.
- Rudell, D. R., Mattheis, J. P., & Hertog, M. L. A. T. M. (2009). Metabolomic change precedes apple superficial scald symptoms. *Journal of Agricultural and Food Chemistry*, 57, 8459–8466.
- Ruth, L. (2003). Tailoring thresholds for GMO testing. Social and economic factors shape new regulations that in turn drive the technology. *Analytical Chemistry*, 9, 392–396.
- Ruttan, V. W. (1992). *Sustainable agriculture and the environment. Perspectives on growth and constraints*. Boulder: Westview Press.
- Sacher, J. A. (1973). Senescence and postharvest physiology. *Annual Review of Plant Physiology*, 24, 197–224.
- Sala, J. M. (1998). Involvement of oxidative stress in chilling injury in cold-stored mandarin fruits. *Postharvest Biology and Technology*, 13, 255–261.
- Sánchez-Martín, J., Heald, J., Kingsdon-Smith, A., Winters, A., Rubiales, D., Sanz, M., et al. (2015). A metabolomic study in oats (*Avena sativa*) highlights a drought tolerance mechanism based upon salicylate signalling pathways and the modulation of carbon, antioxidant and photo-oxidative metabolism. *Plant, Cell & Environment*, 38, 1434–1452.
- Saunders, J., & O’Neill, N. (2004). The characterization of defense responses to fungal infection in alfalfa. *BioControl*, 49, 715–728.
- Schilter, B., & Constable, A. (2002). Regulatory control of genetically modified (GM) foods: Likely developments. *Toxicology Letters*, 127, 341–349.
- Schloss, P. D., & Handelsman, J. (2003). Biotechnological prospects from metagenomics. *Current Opinion in Biotechnology*, 14, 303–310.
- Schmidt, A., & Jäger, K. (1992). Open question about sulfur metabolism in plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, 43, 325–349.
- Schmidt, C. (2004). Metabolomics takes its place as latest up-and-coming “omic” science. *Journal of the National Cancer Institute*, 96, 732–734.
- Shanker, A. K., Maheswari, M., Yadav, S. K., Desai, S., Bhanu, D., Attal, N. B., & Venkateswarlu, B. (2014). Drought stress responses in crops. Drought stress responses in crops. *Functional & Integrative Genomics*, 14, 11–22.
- Shao, H. B., Chu, L. Y., Abdul Jaleel, C., & Zhao, C. X. (2008). Water-deficit stress-induced anatomical changes in higher plants. *Comptes Rendus Biologies*, 331, 215–225.

- Shao, H. B., Guo, Q. J., Chu, L. Y., Zhao, X. N., Su, Z. L., Hu, Y., et al. (2007). Understanding molecular mechanism of higher plant plasticity under abiotic stress. *Colloid Surface B*, *54*, 35–54.
- Sanchez, D. H., Schwabe, F., Erban, A., Udvardi, M. K., & Kopka, J. (2012). Comparative metabolomics of drought acclimation in model and forage legumes. *Plant, Cell & Environment*, *35*, 136–149.
- Sharma, K. K., & Lavanya, M. (2002). Recent developments in transgenics for abiotic stress in legumes of the semi-arid tropics. In: *JIRCAS working report* (pp. 61–73). Available from: <https://core.ac.uk/download/pdf/211011777.pdf>.
- Shimada, N., Akashi, T., Aoki, T., & Ayabe, S. (2000). Induction of isoflavonoid pathway in the model legume *Lotus japonicus*: Molecular characterization of enzymes involved in phytoalexin biosynthesis. *Plant Science (Shannon, Ireland)*, *160*, 37–47.
- Smith, S. E. (2002). Soil microbes and plants – Raising interest, mutual gains. *The New Phytologist*, *156*, 142–144.
- Somerville, C., & Somerville, S. (1999). Plant functional genomics. *Science (New York, N.Y.)*, *285*, 380–383.
- Sousa Silva, M., Cordeiro, C., Roessner, U., & Figueiredo, A. (2019). Editorial: Metabolomics in crop research—Current and emerging methodologies. *Frontiers in Plant Science*, *10*, 1013. Available from <https://doi.org/10.3389/fpls.2019.01013>.
- Sparks, D. L. (1987). Potassium dynamics in soils. In B. A. Stewart (Ed.), *Advances in soil science* (vol. 6, pp. 1–63). New York: Springer.
- Strehmel, N., Praeger, U., König, C., Fehrle, I., Erban, A., & Geyer, M. (2010). Time course effects on primary metabolism of potato (*Solanum tuberosum*) tuber tissue after mechanical impact. *Postharvest Biology and Technology*, *56*, 109–116.
- Subbiah, M. T. R. (2006). Nutrigenetics and nutraceuticals: The next wave riding on personalized medicine. *Translational Research: The Journal of Laboratory and Clinical Medicine*, *149*, 55–61.
- Sun, X., Zhu, A., Liu, S., Sheng, L., Ma, Q., Zhang, L., et al. (2013). Integration of metabolomics and subcellular organelle expression microarray to increase understanding the organic acid changes in post-harvest citrus fruit. *Journal of Integrative Plant Biology*, *55*, 1038–1053.
- Tan, K., Keltjens, W. G., & Findenegg, G. R. (1991). Role of magnesium in combination with liming in alleviating acid-soil stress with the aluminum-sensitive sorghum genotype CV323. *Plant and Soil*, *136*, 65–72.
- Tisdale, S. L., Nelson, W. L., & Beaton, J. D. (1985). *Soil fertility and fertilizers*. New York: MacMillan.
- Trujillo, E., Davis, C., & Milner, J. (2006). Nutrigenomics, proteomics, metabolomics and the practice of diets. *Journal of the American Dietetic Association*, *106*, 403–414.
- UN. (1991). *Guidelines for consumer protection with proposed new elements on sustainable consumption. Agenda 21*. New York: United Nations.
- UNDP. (1998). *Human development report. Report no. 4*. New York: United Nations Program for Development.
- US Senate Committee on Agriculture, Nutrition, & Forestry. 101-624 – *Food, Agriculture, Conservation, and Trade Act of 1990*. (1990). <<https://www.agriculture.senate.gov/imo/media/doc/101-624.pdf>> Accessed 21.01.21.
- Uarrotta, V. G., & Maraschin, M. (2015). Metabolomic, enzymatic, and histochemical analyzes of cassava roots during postharvest physiological deterioration. *BMC Research Notes*, *8*, 648. Available from <https://doi.org/10.1186/s13104-015-1580-1583>.
- Urano, U., Kurihara, Y., Seki, M., & Shinozaki, K. (2010). Omics' analyses of regulatory networks in plant abiotic stress responses. *Current Opinion in Plant Biology*, *13*, 132–138.
- Vance, C. P. (1996). Root-bacteria interactions: Symbiotic N₂ fixation. In Y. Waisel, A. Eshel, & U. Kafkaf (Eds.), *Plant roots: The hidden half* (pp. 839–868). New York: Marcel Dekker.
- Vance, C. P. (1997). Enhanced agricultural sustainability through biological nitrogen fixation. In A. Legocki, H. Bothe, & A. Pühler (Eds.), *Biological fixation of nitrogen for ecology and sustainable agriculture* (pp. 179–186). Berlin: Springer-Verlag.
- Vance, C. P., & Graham, P. H. (1994). Nitrogen fixation in agriculture: Application and perspectives. In A. I. Tikhonovich, N. A. Provorov, V. I. Romanov, & W. E. Newton (Eds.), *Nitrogen fixation: Fundamentals and applications* (pp. 77–86). Dordrecht: Kluwer Academic.
- Van Elsas, D., Wellington, E. M. H., & Trevors, J. T. (1997). *Modern soil microbiology*. New York: Marcel Dekker.
- Vílchez, J. I., Niehaus, K., Dowling, D. N., González-López, J., & Manzanera, M. (2018). Protection of pepper plants from drought by *Microbacterium* sp. 3J1 by modulation of the plant's glutamine and α-ketoglutarate content: A comparative metabolomics approach. *Frontiers in Microbiology*, *9*, 284. Available from <https://doi.org/10.3389/fmicb.2018.00284>.
- Wang, H., Hwang, S. F., Chang, K. F., Turnbull, G. D., & Howard, R. J. (2003). Suppression of important pea diseases by bacterial antagonists. *BioControl*, *48*, 447–460.
- Wang, R., Okamoto, M., Xing, X., & Crawford, N. M. (2003). Microarray analysis of the nitrate response in Arabidopsis roots and shoots reveals over 1,000 rapidly responding genes and new linkages to glucose, trehalose-6-phosphate, iron, and sulfate metabolism. *Plant Physiology*, *132*, 556–567.
- Warkentin, T. D., Rashid, K., & Xue, A. G. (1996). Fungicidal control of ascochyta in field. *Canadian Journal of Plant Science*, *76*, 67–71.
- White, P. J., & Broadley, M. R. (2001). Chloride in soils and its uptake and movement within the plant: A review. *Annals of Botany*, *88*, 967–988.
- WRI. (1998). *World resources: A guide to the global environment. Report*. Washington, DC: World Resources Institute.
- Wu, X. R., Chen, Z. H., & Folk, W. R. (2003). Enrichment of cereal protein lysine content by altered tRNA^{lys} coding during protein synthesis. *Plant Biotechnology Journal*, *1*, 187–194.
- Wu, Q. D., & VanEtten, H. D. (2004). Introduction of plant and fungal genes into pea (*Pisum sativum* L.) hairy roots reduces their ability to produce pisatin and affects their response to a fungal pathogen. *Molecular Plant-Microbe Interactions*, *17*, 798–804.
- Yang, L., Fountain, J. C., Ji, P., Ni, X., Chen, S., Lee, R. D., et al. (2018). Deciphering drought-induced metabolic responses and regulation in developing maize kernels. *Plant Biotechnology Journal*, *16*, 1616–1628.
- Yoshida, M., Cowgill, S. E., & Wightman, J. A. (1997). Roles of oxalic and malic acids in chickpea trichome exudates in host-plant resistant to *Helicoverpa armigera*. *Journal of Chemical Ecology*, *23*, 1195–1210.

- You, J., Zhang, Y., Liu, A., Li, D., Wang, X., Dossa, K., et al. (2019). Transcriptomic and metabolomic profiling of drought-tolerant and susceptible sesame genotypes in response to drought stress. *BMC Plant Biology*, 19, 267. Available from <https://doi.org/10.1186/s12870-019-1880-1>.
- Yun, Z., Qu, H., Wang, H., Zhu, F., Zhang, Z., Duan, X., et al. (2016). Comparative transcriptome and metabolome provides new insights into the regulatory mechanisms of accelerated senescence in litchi fruit after cold storage. *Scientific Reports*, 6, 19356. Available from <https://doi.org/10.1038/srep19356>.
- Zarate, R. (2010). Plant secondary metabolism engineering: Methods, strategies, advances, and omics. *Comp Nat Prod*, 3, 629–668.

Plant metabolomics: a new era in the advancement of agricultural research

Priyanka Narad, Romasha Gupta and Abhishek Sengupta

Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India

9.1 An introduction to metabolomics

The processes that occur during the life cycle include synthesis and breakdown of major biological molecules such as proteins, nucleic acids, and carbohydrates, termed metabolism. Metabolomics can capture the various biochemical, nutritional, and toxicological features of an organism by taking into consideration the metabolite composition which has a close relationship to the organism's phenotype. Thus metabolomics serves as an important aid for the investigation of metabolic composition of the crops that are genetically modified (GM) (Li et al., 2018).

The exploitation of modern metabolomic platforms can explore the regulatory networks and explain the complex biological pathways that are involved in the control of crop growth and development.

Metabolomics is an ultimate tool to understand the complex nature of biological systems. Metabolites in such systems, having molecular weight less than 1500 Da (sometimes 30–3000 Da) can be both identified and quantified (Dunn, Bailey, & Johnson, 2005). Metabolome refers to the group of tiny molecules present in the cell of an organism and consists of various molecules such as amino acids, peptides, carbohydrates, nucleic acids, vitamins, organic acids, flavonoids, alkaloids, polyphenols, or any other compound that is synthesized or metabolized by a cell. Metabolites are essential because they play an important part in the behavior of the individual containing them. Since these products serve as the final products of the regulatory processes of the cell, they are responsible for the responses produced by the biological systems to genetic changes. Due to this, metabolomics is taken as a link between phenotype and genotype (Fiehn, 2002).

The basic idea of metabolomics is to deal with the genetic improvement of crops based on their chemical composition, which may be nutritional or functional aspect, or the activity of chemical compounds in providing resistance to certain plant species. Metabolites are vital components of plant metabolism due to their effect on plant architecture and their biomass. Subsequently, metabolomics has become one of the major breakthroughs in science, flooring a way of accuracy for metabolite profiling in various organisms.

Metabolomics depicts the physiological state of a cell as well as helps to solve the gene's function by depicting the various layers of genes involved in the regulation and interception of metabolic pathways. An integrated approach with various omics studies has allowed the researchers to improve the important traits in crop species by exploring the regulatory steps associated with them such as epigenetic regulation, posttranscriptional, and posttranslational modifications (Parry & Hawkesford, 2012). Metabolomics possesses the potential to select superior traits and improve the breeding materials by utilizing the available whole genome sequence and genome-wide genetic variants by effectively integrating metabolomics in crop breeding programs. There are several methods and tools in metabolomics that are employed for substantial improvement of crops such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy (Farag, 2014). These tools and techniques allow large-scale metabolite surveys that result in a considerable amount of data, supporting in plant improvement schemes.

Nevertheless, the integration of metabolic analysis with various omics data to understand plant development still remains as a major challenge, as the relationship between these omics is complex in nature. Yet plant metabolomics has developed itself as a powerful tool to explore the various aspects of plant physiology and biology, which ultimately

enhance the knowledge of the metabolic and molecular regulatory mechanisms that regulate plant growth, development, stress responses, and the improvement of crop productivity and quality (Schauer & Fernie, 2006).

9.2 Significance of metabolomics in plant biotechnology

Since the mid-1990s, metabolomics has been recognized in the field of plant biology as several studies were reported for gene identification. *Arabidopsis thaliana* is a model plant as it has been the foremost thoroughly researched due to the availability of its extensive resources for studying functional genomics (Benfey et al., 2007).

Metabolomics is counted as one of the most emerging and interesting approaches of omics tools for crop improvement as it decrypts abiotic stress tolerance in plants. Both biotic and abiotic stresses play a significant role in the reduction of crop yield (Atkinson & Urwin, 2012) (Fig. 9.1). Although plants have a similar mechanism to respond to both the types of stresses, yet these stresses produce different variations in plants' physiological as well as biochemical processes.

To impart stress resistance, the plant synthesizes phytohormones at the emergence of abiotic stress. The oxidative stress brings a disturbance in the stomatal conductance of the plant and activates several signaling mechanisms (Robinson, Heath, & Mansfield, 1998). Thus, in a particular plant species, the specific phenomenon of gene expression profile depicts the precise and overall composition of metabolites. As a result, due to the activation of a particular metabolic network, a unique bioactive agent is synthesized (Jamil, Riaz, Ashraf, & Foolad, 2011).

There has been tremendous progress in recent years in the field of metabolomics. However, some bottlenecks got to be particularly addressed to take advantage of metabolomics to its full potential. Once these bottlenecks are removed, new platforms could be explored for crop improvement, which will ultimately guarantee global food security (De Filippis, 2018). As of now, the metabolomic analytical tools are still lacking to detect all the metabolites in sample tissues. This drawback is due to the direct association with the cell's biological modification and complex chemical nature of metabolites in the plant metabolome. To identify the major pitfalls in metabolomics research and to understand appropriately the whole metabolome profile, technical bottlenecks and biological bottlenecks are used for broad range coverage and to draw efficient knowledge (Allwood, Vos, et al., 2011). It is nearly hard to spot the metabolites using the current analytical techniques due to the wide range and diverse chemical composition. To bring precision for a whole metabolite coverage, there has been an advancement in analytical instruments, such as improved NMR. This

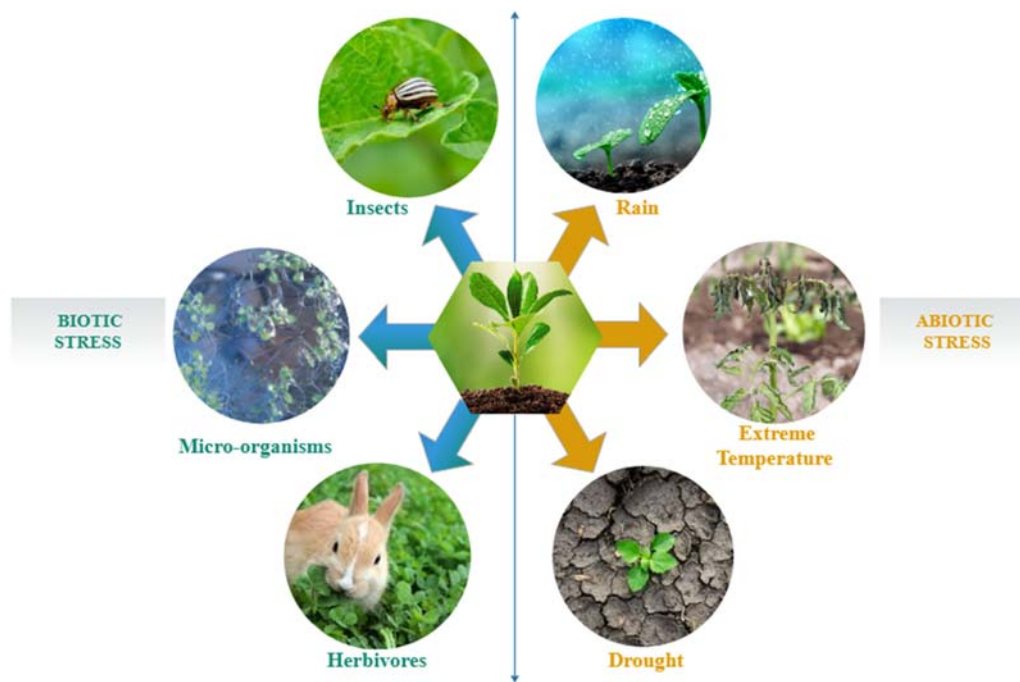


FIGURE 9.1 Diagrammatic representation of biotic/abiotic stress in plants. The causative agents of biotic stress are commonly the living organisms such as insects, microorganisms, bacteria, fungi, viruses that directly derive their nutrients from the host and can lead to death of plants. Abiotic stresses severely affect all essential mechanisms in plants from germination to maturity.

coverage is predicted to improve progressively with the advancement of technology. Still, the major problem in metabolomics lies in the identification and interpretation of a huge number of unexplored plant metabolites (Monteiro, Carvalho, Bastos, & Guedes de Pinho, 2013). Thus it is expected that while designing the advanced tools to beat these hurdles, more sophisticated approaches will be used to explore the accurate features of any metabolite and elucidate its biological function.

9.3 Technologies involved in metabolomics improvement

There has been an advancement in analytical techniques such as MS and NMR spectroscopy with bioinformatics, to study metabolomics. To tackle abiotic/biotic stress in plants, metabolomics tools have been integrated with other omics-based tools such as genomics, proteomics, and transcriptomics (Fukushima, Kusano, Redestig, Arita, & Saito, 2009) (Fig. 9.2).

Two important techniques of modern metabolomics platforms, NMR and MS, involve the generation of metabolome data. The NMR-based metabolite detection depends upon the utilization of magnetic properties of nuclei of atoms under the magnetic field. It is a nondestructive method and is extensively used to identify metabolites having lower molecular weight protein. It reveals the structures of protein–ligand complexes and helps in direct binding of the target protein by retaining its use over MS. The GC (gas chromatography)–MS platform is extensively used for nontargeted analysis (Zhang, Sun, Wang, Han, & Wang, 2012). GC–MS approach includes derivatives of samples and then makes the compounds volatile; due to which the underivative compounds (except hydrocarbon) remain unnoticed during analysis. There has been an improvement in the separation of coeluting peaks (deconvoluted peak) with the introduction of GC X GC–TOF (time-of-flight)–MS, which enhances higher sample throughput. Liquid chromatography (LC)–MS usually uses electrospray ionization (ESI) and atmospheric pressure chemical ionization (APCI) and it has been widely used for both targeted and nontargeted approaches to detect primary and secondary metabolites of higher mass, that is, the exhaustive dataset that are generated from abovementioned high-throughput tools are processed through data processing platforms such as MET-COFEA, Met-XAlign, and ChromaTOF. This includes alignment, baseline correction, separation of coeluting peaks, normalization of data before the identification of compounds (Kim, Ouyang, Jeong, Shen, & Zhang, 2014). For the identification of metabolites, there are many metabolome databases such as NIST, METLIN, and GOLM. The identified metabolites are then further subjected to statistical analyses such as principal component analysis (PCA), K-means clustering, correlation map, heat map, boxplot, and reconstructing metabolic pathways, by using tools and software such as Cytoscape, MetaboAnalyst, and statistical analysis tool (Chong et al., 2018; Shannon et al., 2003). These analyses are used to identify and monitor the metabolic markers related with varied agronomic traits.

Depending on the chemical nature of the compounds, technologies such as NMR, GC, and high-performance LC (HPLC) coupled with MS, as well as capillary electrophoresis (CE) coupled with MS are used in metabolomics (Sato, Soga, Nishioka, & Tomita, 2004). The application of NMR adapted in agriculture can be seen in quality control, analysis of GM plants, chemotaxonomy (classification and characterization), besides the study of diseases in humans. The main

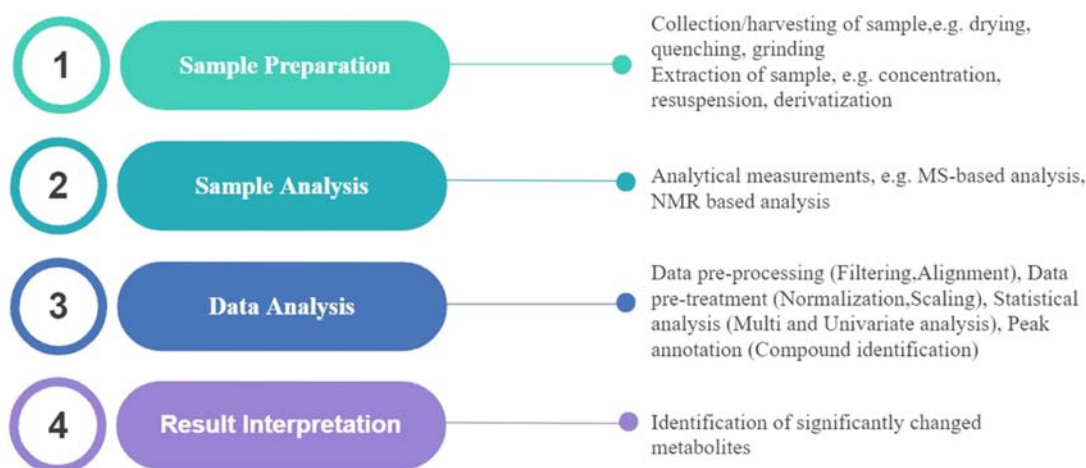


FIGURE 9.2 Flowchart of various steps technologies used in the improvement of plant metabolomics starting from sample preparation with its collection, processing, storage to data acquisition with MS-/NMR-based techniques, data analysis for compound identification and biochemical interpretation. *MS*, Mass spectrometry; *NMR*, nuclear magnetic resonance.

advantages of the NMR as compared to the rest of the analytical techniques are that it can detect a vast range of chemical compounds of distinct nature, identification of metabolite is simple, quantification does not pose any issue and it is highly reproducible. In addition to this, the most important point is that this method is quick and simple and does minimal damage to the existing compounds during the preparation of the extracts (Serkova & Niemann, 2006). However, there is less sensitivity to this technique. NMR is one of the major applications of chemotaxonomy because it is possible to classify and identify plants and their derived preparations through the obtained metabolomics profiles. Examples of the application of NMR in chemotaxonomy are the classification of *Cannabis sativa*, *Ilex* species, *Ephedra*, commercial samples of *Catuaba*, and discrimination of commercial preparations of *Matricaria*.

Despite the technological advancements, it is practically not possible to determine the overall composition of a single cell with a single analytical technique. To do this, coupled techniques are used, such as HPLC and GC coupled with MS (Lopes, Santa Cruz, Sussulini, & Klassen, 2017; Vaclavik, Lacina, Hajslova, & Zweigenbaum, 2011). These analyses are done by using analytical techniques involved in separation, identification, and quantification. These techniques must have high resolution, high accuracy, and very sensitive and be able to analyze a wide range of compounds of various chemical natures and origins. Such specifications are required because the structural complexity of many molecules makes their study difficult.

Currently, we have various analytical techniques that are developing at a fast pace to obtain reliable data about the behavior of a plant species or its response to the diverse climatic factors. Improvements in a wide range of applications for genetic studies can be seen during the last years. The qualitative and quantitative analysis of metabolites is an important aspect of the metabolomic study as it reveals the biochemical state of an organism (Oldiges et al., 2007). This information can then be used to identify and evaluate the function of the gene and the various responses of the organism in the conditions where it develops.

9.4 Metabolomics databases

The rapid development of metabolomic databases has been an aid in the metabolite annotation. Computational informatics has become a prior requirement of metabolomic experiments. Different online web-based programs have been designed during the last few years to aid metabolomics in data assessment, data mining, data processing, and data interpretation (Sugimoto, Kawakami, Robert, Soga, & Tomita, 2012) (Table 9.1). The removal of accurate and monetary assessable platforms has gigantically facilitated the design and maintenance of web tools that can be utilized by many researchers with little bioinformatics knowledge and limited computational facilities.

However, the Internet poses substantial drawbacks while handling the huge raw datasets frequently. An online bioinformatics tool called XCMS allows the uploading of raw data directly and guides the user in statistical analysis and data processing (Tautenhahn, Patti, Rinehart, & Siuzdak, 2012). But XCMS servers fail to deal with the huge data files due to limited space. Recently, the establishment of a XCMS stream for programmed data transfer in LC–MS experiments is done which reduces the data processing time and enhances the efficacy of an online system (Montenegro-Burke et al., 2017). It also helps to detect mutative substances through MS tools by applying the METLIN database (Guijas et al., 2018). To carry out statistical investigation and metabolite detection through the MS/MS database and formula predictor, R-scripts are programmed in-house to get output comprising characteristics in the formation via the XCMS package. Another online database called METLIN is applied in numerous studies related to stress response metabolic profiling in plants. It helps in the metabolic profiling of specific metabolites and allows immediate retrieval of LC/MS, MS/MS, and Fourier transform mass spectrometry (FTMS) analysis outcomes by permitting its operator to put a query in the database through a programmed framework. Another web-based tool known as MetaGeneAlyse is used for the implementation of the regular clustering technique like K-means and independent component analysis. There has been an efficient modification of MetaboAnalyst, by integrating several tools such as MSEA and MetPA. A significant web-based tool called MeltDB has been employed for assessing data, processing them, and then carrying out statistical analysis in plant metabolomics (Neuweger et al., 2008). There are several other databases which do not require any local installation and are operated by windows GUIs (graphical user interfaces) that are iMet-Q, MetAlign, and MS-Dial. MZedDB and KEGG have been extensively used to study the metabolome with a species-specific origin or species-nonspecific origin. A new tool has been developed recently called Galaxy-M, for examining the untargeted metabolites using LC–MS techniques. Meta box is another online server that possesses various uses in the elucidation of omics data. Two extensively used omics-based web tools that are used to perform univariate and multivariate statistical analysis, interpretation of gene expression data, and visualization of metabolomics data are GenePattern and Babelomics. As a result, an integrated analysis of the *Arabidopsis* metabolome, based on the AtMetExpress database, and *Arabidopsis* transcriptome, based on the AtGenExpress database, has allowed a

TABLE 9.1 Various tools used for data analysis in plant metabolomics that provides an invaluable help for researchers to understand the biological mechanisms responsible for the variance in the experimental metabolomic profiles.

S.no	Function	Tools	Weblinks
1	<i>Data processing</i>	MeltDB 2.0	https://meltldb.cebitec.uni-bielefeld.de/cgi-bin/login.cgi
		MetAlign	www.metalign.nl
		MET-COFEA	http://bioinfo.noble.org/manuscript-support/met-cofea/
		iMet-Q	http://ms.iis.sinica.edu.tw/comics/Software_iMet-Q.html
		XCMS	http://bioconductor.org/packages/release/bioc/html/xcms.html
		MAVEN	https://maven.apache.org/
		MZmine2	http://mzmine.github.io/
		MS-Dial	http://prime.psc.riken.jp/compms/msdial/main.html
		MaxQuant	https://www.maxquant.org/
2	<i>Data annotation</i>	MetaboSearch	http://omics.georgetown.edu/metabosearch.html
		MetiTree	http://www.metitree.nl/
		Metacrop 2.0	http://metacrop.ipk-gatersleben.de/apex/f?p = 269:111
		MetAssign	http://mzmatch.sourceforge.net/
		MZedDB	http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html
		MaxQuant	https://www.maxquant.org/
3	<i>Data analysis</i>	metaP-server	http://metap.helmholtz-muenchen.de/metap2/
4	<i>Statistical analysis</i>	MetaboAnalyst	https://www.metaboanalyst.ca/
		MetAlign	www.metalign.nl
		Babelomics 5.0	http://babelomics.bioinfo.cipf.es/
		COVAIN	http://www.univie.ac.at/mosys/software.html
		GenePattern	https://www.genepattern.org/
		Cytoscape	https://cytoscape.org/
		MetaScape	https://metascape.org/gp/index.html#/main/step1
5	<i>Workflow analysis</i>	Galaxy-M	https://github.com/Viant-Metabolomics/Galaxy-M
		Metabox	http://kwanjeeraw.github.io/metabox/
6	<i>Pathway analysis</i>	MetExplore	https://metexplore.toulouse.inrae.fr/index.html/
		MetPA	https://www.metaboanalyst.ca/
		Mummichog	https://shuzhao-li.github.io/mummichog.org/
7	<i>Metabolite annotation</i>	METLIN	http://metlin.scripps.edu
		MetFrag	https://ipb-halle.github.io/MetFrag/
		MassBank	https://massbank.eu/MassBank/Search
		MarVis	http://marvis.gobics.de/
		MMCD	http://mmcd.nmr.fam.wisc.edu/
8	<i>Metabolite identification</i>	CFM-ID	https://cfmid.wishartlab.com/
9	<i>Metabolite data analysis</i>	MetaGeneAlyse	https://metagenealyse.mpimp-golm.mpg.de/

holistic correlation between metabolite accumulation and gene expression, thus allowing the identification of related metabolites and genes. These research strategies that are developed in *Arabidopsis* form the basic structure for the metabolomic investigation in crops.

9.5 Metabolite profiling, identification, and quantification

The terms metabolomics, metabolic, or metabolite profiling are alternatively used to explain three kinds of approaches like targeted metabolomics, semitargeted metabolomics, and untargeted metabolomics (Razzaq, Sadia, Raza, Hameed, & Saleem, 2019). There are certain factors involved as a prerequisite of efficient metabolic profiling like methods for sample preparation, the accuracy of the experiment, quantification and detection of metabolites, and evaluation of specific targets, whereas for an untargeted approach, it is necessary to detect the structural and chemical composition of metabolites while performing targeted and semitargeted studies. This enables the evaluation of the chemical nature of metabolites before data procurement.

With the advancement of technology, metabolite profiling is growing at a very fast rate and has become useful for phenotyping as well as diagnostic analyses of plants (Kopka, Fernie, Weckwerth, Gibon, & Stütt, 2004). It has become a key tool in understanding the cellular response to biological conditions. The aim of improving the compositional quality of crops has always been achieved by metabolomics approaches as they are used to assess the natural variance in metabolite content between specific plants. Metabolite profiling has a huge contribution in various areas.

Metabolomics is species independent, that is, it can be used for a wide variety of species and comparatively requires very little time for reoptimizing protocols for a new species. Earlier metabolite profiling made use of metabolite composition as a diagnostic tool to establish the equivalence of GM and conventional crops, the metabolic response to herbicide, and the classification of plant genotypes. Recently, it is being abundantly employed in describing the response of plants to a wide range of biotic or abiotic stresses (Piasecka, Kachlicki, & Stobiecki, 2019). Metabolite profiling is also extensively used in deciphering gene function, investigating the metabolic regulation and analyzing the systemic response to environmental or genetic perturbations. According to a recently published report in this area, it is seen that the combination of metabolite profiling and marker-assisted selection proves extremely informative in a better understanding of the chemical composition of crop species.

One of the major challenges in metabolomics is the identification of metabolites. It becomes difficult to identify the chemical and physical diversities of metabolites based on MS data. Currently, in untargeted metabolomics analysis, the identification of metabolite is majorly done by a mass-based search which is followed by manual verification. The first step is to search the m/z value of a molecular ion of interest against a database. Then there is retrieval of the metabolites bearing molecular weights within a specified tolerance range to the query m/z value from the databases as putative identifications. The actual compounds of these putative identifications are exposed to a tandem MS (MS/MS) experiment along with the sample (Wang et al., 2017) (Fig. 9.3). Through the comparisons made in sample between the MS/MS spectra and retention times of the authentic compounds with the molecules of interest, the identity of the molecules is confirmed. Due to the existence of isomers and the restricted accuracy of mass spectrometers, putative identifications from mass-based searches are rarely unique. At times, a certain molecule ion can have more than 100 putative identifications, thus making the manual verification expensive and difficult. Therefore this method is only intended for a limited number of molecules. A computational framework is rather suggested to improve the productivity of metabolite identification for an enormous number of metabolites as it can decrease the number of putative identifications and prioritize them.

Quantitation of metabolites is one of the aims in metabolomics for the evaluation of changes occurring in response to disease, treatment, environmental, and genetic disturbances. The use of QqQ-based LC–SRM (selected reaction monitoring)—MS/MS has turned out to be of absolute choice for targeted metabolomics studies (Xiao, Zhou, & Resson, 2012). Due to the diverse chemical properties of metabolites and the magnitude difference in their concentration, it becomes one of the greatest challenges in such studies. So, it is nearly impossible to quantify all the metabolites at once on any platform. However, metabolite quantitation is facilitated by coupling LC to MS, yet there is no single LC method that may prove to be ideal for the separation of all classes of metabolites. Thus to detect a wider range of metabolites from different biological samples, many efforts are being made to improve LC-separation capacity.

9.6 Metabolic engineering in plants

The framework analysis of networks for metabolite accumulation or gene expression has increased the understanding of cellular processes and response of cells to biological perturbations (Ideker et al., 2001).

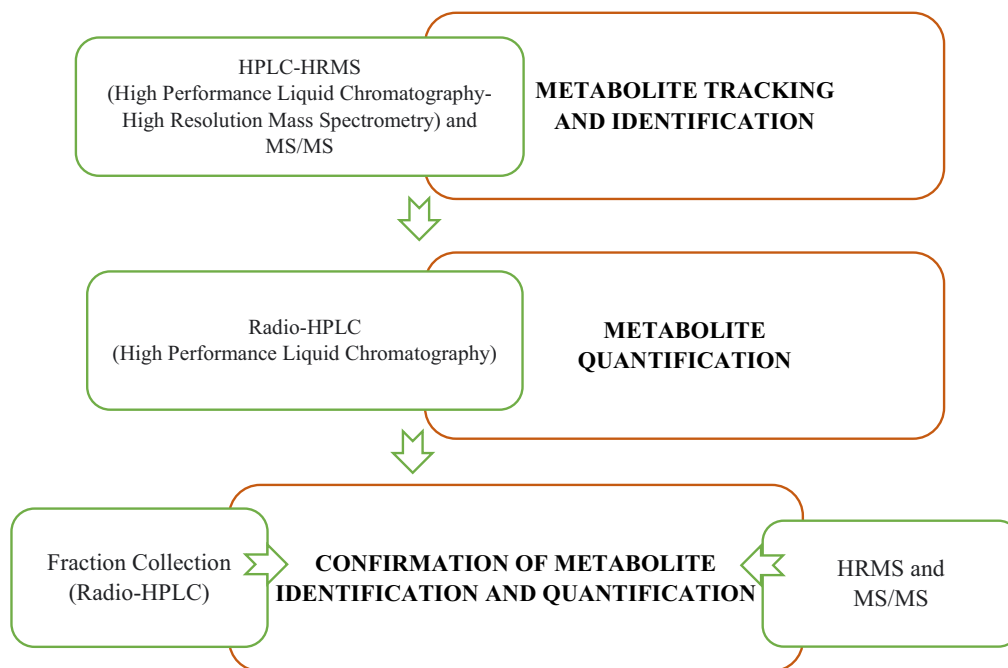


FIGURE 9.3 Schematic representation of metabolite identification and quantification techniques to observe various metabolic activities of metabolites containing a wide variety of physicochemical properties and to perform high-throughput analysis.

Rice is an important cereal grain across worldwide by providing nutrients for both humans and animals, is found to encode ~32,000 genes in its genome. Yet the biological functions of more than half of these genes are still unknown. The basic approach used to identify the novel genes in rice was gain- and loss-of-function approaches. To investigate the direct relationships among metabolic composition, genotypes, and phenotypes for agronomical traits, association analyses with the genetic core collections and segregating populations were employed. There is a varied usage of plant species that are destined for agriculture, ranging from traditional foods to those possessing nutritional value, desirable traits, and the different industrial products produced from them such as fibers, polymers, packaging materials, basic chemical building blocks, and fuels (Sanderson et al., 2004). The basic aim of metabolomics tools in agriculture is to identify the biochemistry and the functions of the species involved to use that knowledge for food and environmental security as well as to utilize their potential in the improvement of nutrition, health and diets, to use them in plants for genetic improvement by considering some of their remarkable characteristics. Due to the continuous rise of population centers, there is a rapid increase in food demand which in turn is increasing the demand for productivity and diversity of basic crops. Trials and studies of genetic improvement of crops are being conducted currently to increase the quantity and quality of yield by avoiding damages caused by pests and by developing resistance to several factors, especially environmental ones. The major crops involved in such trials are potato, rice, tomato, maize, etc.

The interaction between plant and pathogen is a unique feature in metabolomic engineering since plants are often influenced by environmental factors (Bino et al., 2004). These factors make the plants to adapt and change parts of their functioning such that they can protect themselves in most cases. One such interaction is the existence of plants with microorganisms, which causes physiological and developmental changes in plants. An example is the interaction of nitrogen-fixing bacteria with legumes. A visible change that occurs in plants during their interaction with a pathogen is the production of various types of compounds that act as a type of defense repellents, attractants, feeding inhibitors or the production of compounds that prove to be beneficial for human health. This provides an overview of the whole system of the biochemical and physiological changes that occurs during the interaction. In a study conducted by the researchers on plant–fungus interaction, in which MS using ESI was used to detect changes in the levels of lipids and hormones, it was predicted that these molecules were involved in the interaction between *Brachypodium distachyon* and *Magnaporthe grisea* (William Allwood, Ellis, Heald, Goodacre, & Mur, 2006). The main response of the plant to the attack of the fungus was detected by a variation in the level of phospholipid. Both targeted and nontargeted studies can be performed at the same time by using metabolomics, for example, the interaction between *Lupinus angustifolius* with the fungus *Colletotrichum lupini*.

By altering photosynthate levels the metabolic composition of fruit is significantly correlated with the fruit dimensions, development, and weight. Gradually as the fruit develops, a substantial change of organic and sugars is seen that determines the final quality of the ripe fruits. The development of fruits depends on the translocation of photoassimilates of leaves than that of own photosynthesis products by acting as a sink. For example, the incubation of tomato plant significantly reduces fruit size and shape in the dark due to the repression of cell cycle genes of fruit which severely affects the cell number and cell (Gonzalez, Gévaudant, Hernould, Chevalier, & Mouras, 2007).

The evaluation of metabolite responses to stress has been geared up by the response of MS-based plant metabolomics. For instance, the stress response hormone abscisic acid (ABA) signals shoot for antitranspiration activities like stomatal closure during water deficit condition, reduction of leaf size, and facilitation of deeper root growth by changing the root architecture under a scarcity of water and nitrogen deficiency (Jackson, 1997).

Lack of vitamin A causes night blindness and in the long run, it might also lead to complete blindness. To reduce the deficiency of vitamin A, β -carotene was targeted and used as provitamin A. The initiation of β -carotene as a supplement of vitamin A was initiated by enrichment of rice endosperm to produce golden rice (Datta et al., 2007). This approach includes the upregulation of carotenoid biosynthetic pathways in rice endosperm and transgene expression of phytoene synthase and phytoene desaturase.

The increasing demand for biofuel demand burgeoning petroleum across worldwide has motivated the researchers in metabolomics to explore renewable and alternative sources, such as biodiesel (Pandey, Venkata Mohan, Chang, Hallenbeck, & Larroche, 2019). Currently a better understanding of biochemical pathways is being used to genetically improve biodiesel crop species such as jatropha, pongamia, soybean, and mustard. The quality of the plant is determined by the composition of the oil. *Jatropha curcas* is being grown as an alternative source of energy as the oil content of its seeds is highly rich in polyunsaturated fatty acid mainly linoleic acid which harms the quality and is vulnerable to oxidation. Seed oil as biodiesel has been extended to several plant species such as cotton, sunflower, and mustard. Fermentation of sugars yielding alcohol such as ethanol and butanol also produces biofuels. As plants bear the tendency of designing and producing multifarious chemical compounds that provide human beings as foods and medicines, in plants associated with modern biotechnology will bring more benefits to mankind by effective engineering of metabolic pathways.

9.7 Environmental and ecological metabolomics

The synthesis of a large number of metabolites under different environmental conditions is highly involved in plant growth and development. Environmental metabolomics involves the characterization of the relationship of plants with their environment. It deals with the definite assessment of metabolite levels under a particular plant environment to pinpoint the effects on plant transformation and any changes in their genetic architecture. The application of metabolomics in the environmental sciences is in a high rise, such as understanding organismal reactions to abiotic stressors, including factors such as temperature and pollution, understanding the biotic–biotic interactions such as infection and herbivory. It also marks the development of biomarkers and risk assessment of toxicant exposure, as well as disease diagnosis and monitoring. In a study conducted by Viant the application of metabolomics in the aquatic organism was done (Bundy, Davey, & Viant, 2009). He measured the metabolites and their variability along with the genotypic and phenotypic interpretation. Similarly, Samuelsson and Larsson researched metabolomics in fish (Asakura, Sakata, Yoshida, Date, & Kikuchi, 2014). Moreover, the recent example of metabolite fingerprinting has been an aid to study population dynamics as it helps in identifying the origin of coffee beans by Choi, Choi, Park, Lim, and Kwon (2010), in identifying the populations of tobacco plants from China and Zimbabwe by Li et al. (2011) and in the identification of the populations of the plant *Arabidopsis lyrata* ssp. *petraea* from secluded regions across Europe. Recent research conducted by Scherling et al. (2010) recommends that the competitive ability and subsequent biodiversity of plants are due to the variation in the metabolome within experimental plant communities. It was found that in small herbaceous species compared to taller, there is higher metabolic diversity and the metabolic profiles showed that the amount of carbon and nitrogen was limited in smaller plants when exposed to higher diversity. Later, Field and Lake (2011) demonstrated direct linkage of metabolic diversity with the genotypic plethora within populations of wild plants.

Ecological metabolomics aims to analyze the plant biochemical connections across distinct temporal and spatial systems. It deciphers the conceivable effect of abiotic/biotic stresses on any indispensable biochemical process through metabolite identification in response to environmental factors. It also explains the biochemical nature of numerous significant ecological phenomena such as the effects of parasite load disease occurrence and infection. It also provides an evaluation of the interaction among two trophic levels or numerous effects of abiotic factors with intra- and interspecific linkage. The phenotypic and physiological feedbacks of plants to environmental fluctuations can be explained by

the variations in the concentration of numerous metabolites that gives mechanistic indications for biochemical networks. One such ecometabolomic application is the screening and quantification of the changes that occur in sap constituents under extended drought conditions by Alvarez, Marsh, Schroeder, and Schachtman (2008). In addition to providing insights into the range of compounds in sap, they have also shown that changes in the composition may lead to alterations in signaling and development during drought. Regardless, ecometabolomic applications are not just restricted to the ecophysiology of species.

Consequently, metabolomics gives a quick and sensitive indicator of ecosystem health by permitting the examination of complex ecological systems. Nevertheless, the full potential of ecological metabolomics is yet to be explored.

9.8 Extraction methods in metabolomics

Various extraction strategies are used to extract and segregate the compounds of interest; however, it is essential to remember to utilize a simple method, low consume time, robust, repeatable, and low cost. A wide class of metabolites can be extracted through conventional extraction techniques such as percolation, maceration, Soxhlet extraction, steam refining, or hydrodistillation as they are of low cost, simple, repeatable, and can be utilized for raw plant extraction. Depending on the source of plant material, the amount to be processed relies on the source of plant material which can range from a few grams to higher amounts, but they are tedious. Currently, they are supplemented with modern techniques such as microwaves, ultrasonication, and supercritical fluid extraction (Gupta, Naraniwal, & Kothari, 2012) (Fig. 9.4).

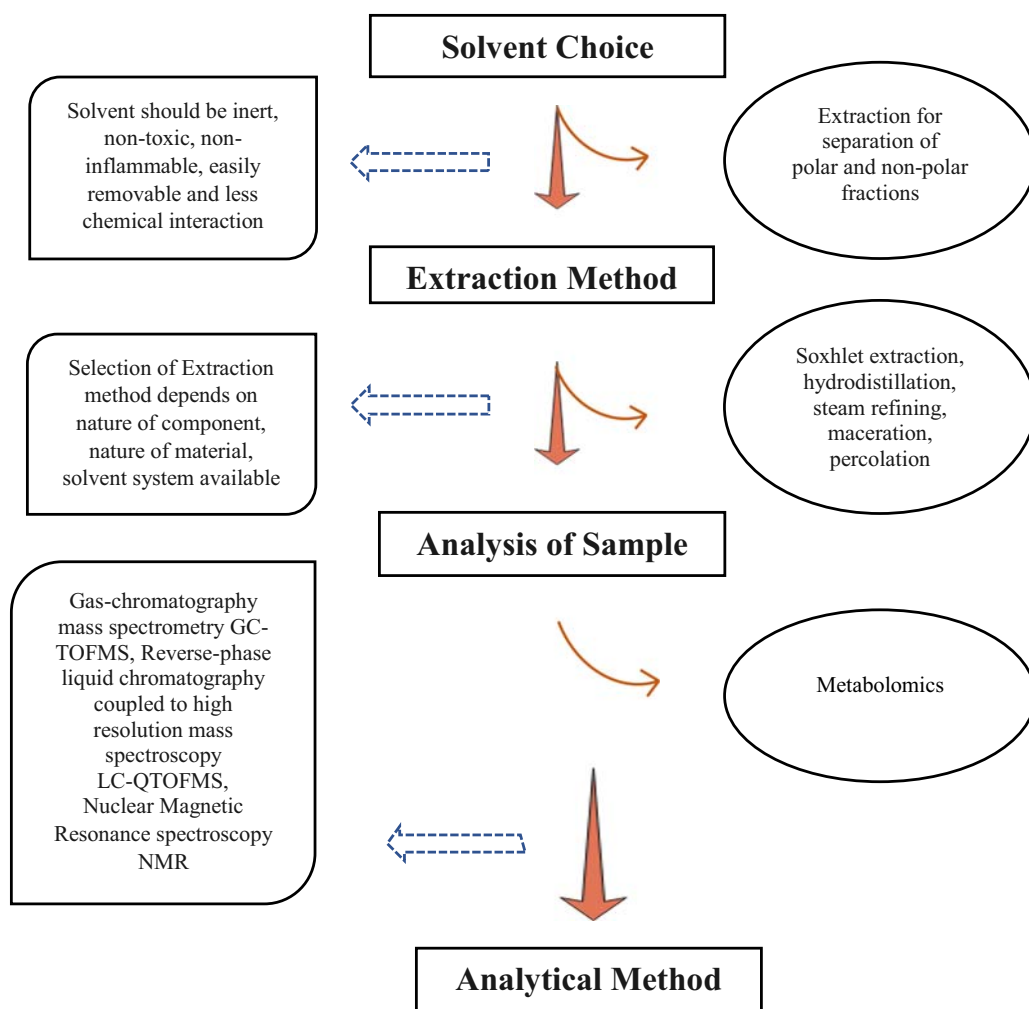


FIGURE 9.4 Stepwise representation of extraction methods in plant metabolomics, that is, preextraction techniques and choice of extraction solvents to increase the efficacy and reproducibility of the sample and to ensure the reliability of a metabolomic study.

No doubt these methods are simple, repeatable, and predominantly with lesser extraction time; however, the expense of these equipment is high. It is likewise imperative to consider the impact of temperature given by the selected method since certain parts of the sample can be disintegrated if the temperature is high. In a recent review done by Roesnner and Dias (De Livera et al., 2012), they marked all the details that ought to produce a better result with the extraction and isolation of the compounds of interest. As soon as the sample is extracted properly, it is prepared to be submitted for the analysis on LC-MS, GC-MS, NMR, or MS.

9.9 Metabolomics-assisted breeding techniques

Remarkable development is seen in the field of metabolomics during the last decade in both instrumentation advancement and software tools' design, giving an excellent chance to check the entire metabolome of different plant species in a high-throughput way. Metabolomic applications have helped various research areas, particularly biotechnology, disease diagnostics and functional genomics, also marking its way for translational metabolomics in plant breeding (Kaddurah-Daouk, Kristal, & Weinshilboum, 2008). The screening process has been accelerated due to the recent advances in postgenomic approaches, also the time required to develop elite crop varieties with improved tolerance against abiotic and biotic stresses has shortened because of the amalgamation of metabolomics with other high-throughput tools. Metabolomics can give a holistic view of various metabolites' diagnosis and phenotyping of plants. Around 840 metabolite units have been recognized in 524 rice cultivars (Wei, Wang, Li, Qu, & Jia, 2018). They have the potential for exploitation in future crop breeding strategies. Availability of the integrated datasets of proteomics, transcriptomics, and metabolomics for mapping the quantitative traits and dissecting genetic variations at the mRNA, protein, and metabolic levels has led the researchers to apply these methods in proteomic quantitative trait locus (pQTL), epigenomic QTL, and metabolic QTL (mQTL) (Jansen, Tesson, Fu, Yang, & McIntyre, 2009) (Fig. 9.5). Genome-wide association studies (GWASs) together with metabolomic techniques (mGWAS) and mQTLs serve as powerful tools for the detection of genetic variations connected with metabolic traits in plants (Luo, 2015).

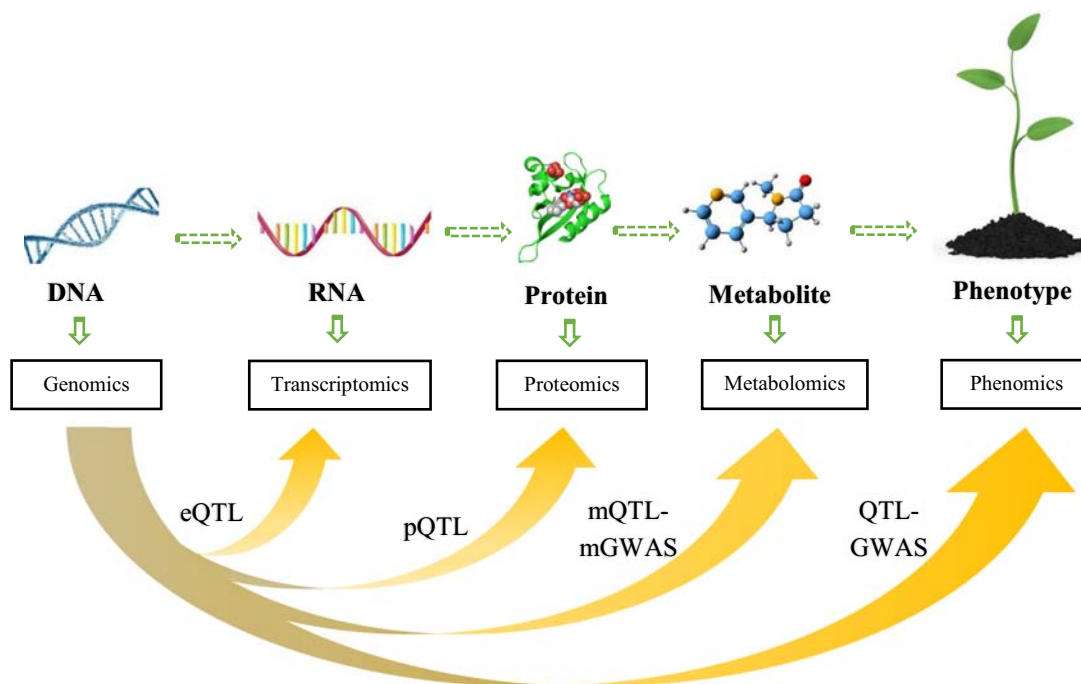


FIGURE 9.5 Genetic association analysis for mapping the gene expression of a metabolite phenotype through QTL mapping and GWAS techniques to overcome the problems arising from different environmental conditions and investigate the effects of genetic deviations on metabolites. Although mGWAS are independent on genetic data, yet MWASSs are dependent on genetic information. *eQTL*, epigenomic QTL; GWAS, genome-wide association; *pQTL*, proteomic QTL; *mQTL*, metabolomic QTL; *mGWAS*, metabolome genome-wide association studies; *QTL*, quantitative trait loci.

9.9.1 Metabolic quantitative trait loci

The potential to use metabolomics in crop breeding is because a metabolome can represent the ultimate phenotype of a cell. It becomes easier to identify and gather information about the target of genomic regions or the genes related to the breeding of crops due to mQTL analysis (Matsuda et al., 2012). Metabolomics-assisted breeding helps the researchers to develop an elite crop of better quality and yield. mQTL investigations provide more comprehensive information about quantitative genetics. The gap between the genotype and phenotype has narrowed down through metabolic profiling as it provides new horizons for metabolic dissecting, starting with the identification of single-nucleotide polymorphism (SNP) markers or candidate gene detection by mQTL mapping analysis. The mQTL approach builds up a linkage between the phenotype and genotype by featuring significant insights into the genetic structure and by analyzing the phenotypic variations through integrated analysis of gene expression and metabolic profiles. Advancement in next-generation sequencing (NGS) has permitted mQTL identifications for candidate genes through ultrahigh-density maps. Biosynthesis of secondary metabolites that are controlled by candidate genes can be detected by utilizing multiomics integrated with forward and reverse genetic approaches (Beleggia et al., 2016). Moreover, population genetics in combination with quantitative genetics and metabolic profiling has begun to reveal hereditary control of the entire metabolome in plants. For example, utilizing a high-density map comprising 1619 bins produced by sequencing, rice mQTL analysis has been performed (Gong et al., 2013). Numerous mQTLs have been detected in flag leaf and germinating seeds across 12 chromosomes. When mQTL analysis was performed for comparative investigations of two rice cultivars, it uncovered the accumulation of tissue-specific secondary metabolites that were under strict genetic control. A sum of 19 metabolites was recognized on 23 loci, suggesting a significant intersection of genetic control in various cells. Similarly, comparative results have been reported for potato, maize, and tomato. The mQTL analysis of backcrossed inbred lines of rice cultivars highlighted 700 various metabolic features. The study uncovered 802 mQTLs having an irregular distribution, which may control distinctive metabolic characteristics. Barley mQTLs, physiological, morphological, and metabolic adaptation was investigated by Templer and colleagues under drought stress conditions. Nearly 57 metabolites and some unique mQTLs, for example, succinate, γ -tocopherol, and succinate were recognized in flag leaf via association genetics. The outcomes demonstrated a molecular basis for barley breeding with expanded resistance against drought stress. Identification of 679 secondary mQTLs of tomato which were linked with environmental stress tolerance was done through the dissection of genomic regions linked with the synthesis of secondary metabolites in wild and introgression lines (Alseekh et al., 2015).

For the identification of traits associated with stress susceptibility, mQTL mapping is an efficient tool. Metabolomics profiling of 179 doubled haploid wheat lines mediated by the LC/MS probed about 558 secondary metabolites, including alkaloids, flavonoids, and phenylpropanoids (Hill et al., 2015). GC-TOF/MS-intervened metabolic analysis of tomato recombinant inbred lines was done to profile seeds and interpret the communication between seed environment, metabolism, and genetics. This study explored many genetic regions that help in the regulation of a set of metabolites. Besides this, many studies distinguished mQTLs controlling the biotic communications in plants. With the advancement in sequencing technology, numerous plant genomes have been sequenced with incessant utilization of mQTL analysis in crop plants.

9.9.2 Metabolic genome-wide association studies

The emergence of mGWAS has proved to be an incredible asset to portray the natural genetic basis of different metabolic changes in the plant metabolome. The global perspective on secondary plant metabolites related to a particular trait has been revealed in a recent study (Hadacek, 2002). Different kinds of flavone glycosylation are being identified in rice varieties through studies made on metabolic polymorphism and revealed a positive correlation of plant growing conditions with introduction to ultraviolet B (UV-B) light. A sum of 175 rice accessions was exposed to metabolomics-assisted GWAS analysis. Identification of 323 associations among 89 secondary metabolites and 143 SNPs were done which indicated two kinds of genetic architecture identifying secondary metabolite concentrations. It is seen that gene-to-metabolic investigation via mGWAS provides a valuable technology for improvement in crop genetics (Dong et al., 2015). The mGWAS analysis has been performed in rice to analyze biochemical and hereditary varieties in its metabolism. It was found that the 36 genes were related to unique metabolites that were responsible for controlling the nourishment and physiological traits. Moreover, five qualities were described, which included a glucosyltransferase, a methyltransferase, and three putative acyltransferases (Chen et al., 2014). The qualities of essential and auxiliary metabolite primary and secondary metabolites can be utilized as metabolic markers to encourage crop breeding for genetic improvement. Understanding the functional genomics in association with plant development, the significance of

advanced metabolomic tools, along with QTL analysis, GWAS and knockout/down technology has been progressively perceived inside the plant science community. Since QTLs are distributed in various regions of the chromosome and a huge count of alleles occur in the process of domestication while searching for candidate genes in correlation with metabolic phenotype in genetic variation in plants, molecular breeding benefits from the fragment with prevailing genes that ultimately leads to high productivity or quality. Although plants are highly suitable for linkage analysis and high-throughput plant phenotyping platforms with respective plant phenomics have offered and integrated a set of novel technologies, yet more detailed information of complex plant phenotypes is still needed to be mined.

The development of NGS technologies has been highly profitable. To understand the genetic mechanisms underlying metabolic diversity and their relationship with complex traits in plants, metabolome-based GWAS (mGWAS) has been used across the world. Even though mGWAS recognizes enormous scope metabolite-related QTL, which can probably be utilized in the future in plants, a few disadvantages are also inescapable at present. Initially, due to the limitation in the present statistical algorithm, it is hard to precisely distinguish the epistasis or gene–environment interaction (G–E) QTL. Second, it is unreasonable for the entire potential genes from one single analysis to be confirmed by transgenic analysis due to the limitation in precision particularly in some regions of the chromosome with the slow decay of linkage disequilibrium and the work and tedious method. Luckily, equivalent to different attributes like seed quality, while the regions of interesting QTL are determined, these QTL could be additionally used for marker-assisted selection breeding without the essentiality to discover the fundamental genes (Pourmortazavi & Hajimirsadeghi, 2007).

9.10 Metabolites present in plant metabolome

Plants produce huge quantities of metabolites that possess diversified structures and abundance. They play significant functions in plant development, growth, and their response to environments. These diverse metabolites having small molecular weight serve not only as the chemical base of crop yield and quality but also as valuable nutrition and sources of energy for human beings and live stocks. The metabolites are generally classified into primary and secondary metabolites. Primary metabolites are the basic requirement for the growth and development of a plant whereas secondary metabolites are significant as they maintain a delicate balance with the environment for a plant to survive under stress conditions (Zaynab et al., 2019).

Primary metabolites in plants are important for the biosynthesis of amino acids, lipids, sugars and are highly conserved in their structures and abundances. During photosynthesis they mediate the glycolysis and tricarboxylic acid cycle, thus affecting the plant growth and development. Variations in the amalgamation of primary metabolites may cause malfunctioning during photosynthesis and imbalanced osmotic adjustment in plants. Varieties in the amalgamation of essential metabolites may prompt photosynthesis breaking down and imbalanced osmotic change in plants. Essential digestion brings about the creation of auxiliary metabolites, similar to flavonoids, atropine, carotenoids, and phytic corrosive. These are not basic for plant endurance and are delivered in light of various pressure conditions, for example, high temperature, chilling, dry season, saltiness, and creepy crawly/bug assault. Production of secondary metabolites such as atropine, flavonoids, phytic acid, and carotenoids is the result of primary metabolism. These secondary metabolites are not critically essential for the survival of the plant and are produced due to different stress conditions like drought, high temperature, salinity, chilling, and insect/pest attack. Secondary metabolites differ widely across plant kingdoms. They may include antioxidants, reactive oxygen species and coenzymes (Dawid & Hille, 2018). Some specialized secondary metabolites may also be present in the plant metabolome which consists of terpenoids (> 25,000), phenolics (~ 10,000), and alkaloids (~ 21,000) that provide tolerance against biotic/abiotic stresses. Some of these specialized compounds have been identified as unique biomarkers that help in measuring plant performance during stress conditions and also serve as essential for crop improvement programs. Primary and secondary metabolites are continuously synthesized through complex biochemical reactions during plant ontogenesis. Thus it is important to reveal the unique metabolic biochemical processes involved in plant biology.

The necessity to explore the underlying biochemical nature is due to the diversity of plant metabolites and their complicated regulatory mechanism. The yield of plant metabolomics relies to a great extent upon its methodologies and instrumentations to extensively distinguish, measure, and localize every metabolite. It is very tough to do so because of the complex nature of these diverse metabolic characteristics and abundances of molecules. Regardless of the fact that accurate and exhausting analysis of the entire metabolome of a biological sample appears to be currently impossible, methodologies and instrumentations of plant metabolomics have been developing at a quick pace to solve this issue (Hegeman, 2010).

Evaluation of food and agronomical traits of crops, especially those of GM crops and their derived GM foods, could be performed in terms of metabolites present. The plant kingdom contains nearly 200,000 compounds of a huge diversity of metabolites and most of them are still unknown. It is estimated that approximately 10,000 secondary metabolites

have been found in various plant species and these discovered metabolites being highly significant plant biology are structurally different in their biochemical properties as well as functions. Metabolomics research is essentially based on low molecular weight metabolites within biological systems and is solely concerned with the identification and quantification of small molecules. Extensive knowledge of biochemical processes that occur during plant metabolism can be known through the metabolic profiling of primary and secondary metabolites. Various methods have been formulated for the detection and identification of specific metabolites. However, no single metabolomics tool can be used for the entire metabolome profiling due to the complex nature of metabolites, massive production in cellular compartments, and diverse chemical composition.

9.11 Workflow of metabolomics analysis

Metabolic analysis involves three core steps for its experimental design, and they are sample preparation, data acquisition through analytical strategies, and utilization of suitable chemo-metric techniques for data mining (Kim & Verpoorte, 2010).

9.11.1 Sample preparation

The most important part of metabolomics is sample preparation as it has a tremendous effect on the final results. For sample material, plant tissues that are above the ground such as stems, seeds, and roots can be utilized (Fig. 9.6). The high-resolution magic-angle spinning technique is broadly used in plant metabolomics tests, even though it isn't suitable for the extraction of plant secondary metabolites that play a crucial role in plants' self-defense mechanism. The basic objective of sample preparation is to enrich the desired metabolites by separating the metabolites from unwanted elements. The best sample preparation technique ought to be fast, efficient, simple, economical, and maintain the sample integrity (Causon & Hann, 2016). Four steps are involved for plant sample preparation for metabolic analysis, harvesting the plant material, quenching, sample extraction, and sample analysis. Depending upon the choice of analytical methods and the characteristics of the metabolites, the extraction and freezing steps can be discarded. Also, quenching of the sample material depends on the biological nature of the sample because harvesting and quenching of the sample material are nearly the same for all analytical tools. High caution should be taken while harvesting the sample as the plant metabolome is delicate to enzymatic reactions that degrade different metabolites. Usually soon after, harvesting the plant material is quenched in liquid nitrogen to avoid any metabolic changes. Another important factor is the age of the plant sample as the metabolic profiling of young leaves is quite different from mature leaves. For sample preparation, it is very critical to avoid enzymatic degradation of the sample material (Harbourne, Marete, Jacquier, & O'Riordan, 2009). Numerous extraction protocols have been created over the last couple of years for metabolomics analysis. Earlier a pestle and mortar are used for grinding leaves but now methods like tissue lyser, electric grinder, and ultrasonic oscillator are used. In metabolite extraction the selection of extraction solvent is also of utmost importance. The solvent should be effortlessly isolated without triggering any biochemical reaction. Some commonly used extraction solvents are aqueous methanol, acetonitrile, ethanol, perchloric acid, and water. The rate of dissolution and solubility highly matters for the choice of extraction protocol. Biological components like cellulose or lignin may collaborate with metabolites and thus influence the dissolution rate. Soxhlet extraction is one of the conventional methods utilized for sample extraction (De Castro & Priego-Capote, 2010). In this technique, continuous heating of the sample is done and concentrated solvent is being used for extraction. For targeted and untargeted metabolic profiling via MS approaches, solid-phase microextraction (SPME) is done. A technique called laser microdissection is used for the isolation of the desired cells from microscopic samples as it does not affect the chemistry and morphology of the desired metabolites in the samples. Another high-speed and accurate method of sample extraction in metabolomics is microwave-assisted extraction. An efficient technique called supercritical fluid extraction can be utilized for volatile

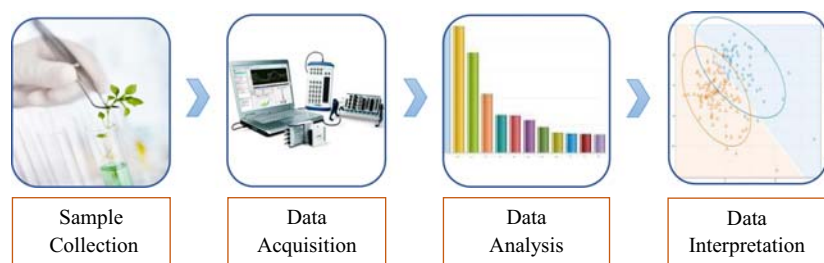


FIGURE 9.6 Workflow of metabolomic analysis in plants involves: sample collection, the type of sample, its storage, preservation and preparation; data acquisition with different techniques and quality control; data analysis, including normalization and identification of metabolites; and data interpretation to handle the high-throughput metabolomics data.

metabolites. Numerous other sample preparation methods exist such as SPME, ultrasound-assisted extraction, enzyme-assisted extraction, and the Swiss rolling technique.

9.11.2 Data mining, annotation, and processing in metabolomics

A deep insight into the molecular complexity of downstream of the genome, proteome, and transcriptome of plants either in normal growth conditions or in response to various stresses has been revolutionized in today's era. An enormous amount of dataset has been established by the whole metabolome analysis due to the huge number of metabolites present in different parts of plant cells or tissues. But the metabolomics data analysis has become more complex because of the complicated nature and composition of metabolites in different plant samples. The main aim of the entire metabolome analysis is to sort the various metabolites of different plant samples initiated by various factors (Aoki-Kinoshita, 2006). Since a considerable amount of data can be created by metabolomics, so it is known as the data-rich technique. Powerful automated tools are required to manage huge datasets and for annotating and storing the raw data. The tough part that rests in plant metabolomics is in extracting accurate information about specific metabolites from massive datasets generated by advanced techniques. The basic steps in data mining are preprocessing, pretreatment, and statistical analysis of data (Liland, 2011; Sun & Weckwerth, 2012). A series of statistical analyses are carried out for the raw data acquired from sample analysis analyses to generate a numerical data matrix and align this data for further processing.

9.11.3 Statistical tools and biomarker identification

Metabolomics can be used to measure the metabolite abundance as a predictive biomarker for disease diagnosis as well as to score the genetic as well as environmental-induced changes in metabolites' concentration. The identification of biomarkers highly depends on the analysis of data using different statistical methods. To estimate the relationship between metabolites and phenotypic variables a proper, multidimensional statistical platform is required for fast forward analysis. A pairwise Pearson's correlation can be utilized to find a particular biomarker, where just a single metabolite is associated with the ideal phenotype. Even though more than one metabolite analysis is needed to design a predictive model, and canonical correlation analysis is often applied to study the maximize correlation between variables (Song, Schreier, Ramírez, & Hasija, 2016). To handle the high-throughput metabolomics data that are mainly adapted from previously existing omics technologies, many statistical tools including those that were originally developed for transcriptomic analysis can be used for metabolomics data analysis. Conventionally, the main aim in any metabolomic analysis is to see the groupwise differences either in a univariate or multivariate technique. Biomarker discovery for univariate analysis is generally performed at initial levels of systems biology, which studies one variable at a specific time. Moreover, it can also confirm the presentation and authenticity of an assumed metabolic marker, whereas multivariate analysis can be utilized for screening plant cultivars and ecotypes, metabolic marker discovery, and disease diagnosis. These tools help to effectively compare the evaluation among various genotypes and samples. Several multivariate statistical tools are available such as PCA, ANOVA (analysis of variance), analysis of variance-simultaneous component analysis (A-SCA), partial least squares-discriminant analysis (PLS-DA), and heat map analysis (Ren, Hinzman, Kang, Szczesniak, & Lu, 2015). Multivariate statistical strategies are generally categorized into two approaches to study high-throughput metabolomics data (Weckwerth & Morgenthal, 2005): unsupervised approach in which unidentified samples are statistically analyzed, keeping the main focus on the natural structure that exists in a dataset, and supervised approach, in which the basic aim is to alter multivariate datasets from metabolic analysis to the demonstrations of biological units under supervision. The supervised technique explains the relationship between the input and output observable in a particular sample of data. One of the most crucial unsupervised multivariate statistical tools that are being extensively utilized for the multidimensional reduction approach is PCA, which is beneficial and efficient because the difference among the various samples can be divided and comprehensively explained in numerous principal components (Xu & Goodacre, 2012), although PCA can't separate variance in samples whenever a multipurpose factor is strongly connected. So, to manage noisy and highly collinear datasets, the PLS approach is used as its extensions, like orthogonal PLS (OPLS), PLS-DA, and sparse PLS, are often frequently performed in metabolic data analysis. OPLS and PLS techniques give huge data that can be helpful for metabolic marker selection. Commercially accessible statistical tools that propose several types of procedures are Matlab and SIMCA-P. There are some excellent R programming software that are developed for different applications in plant metabolomics research. The R package language statistical tools provide statistical graphics and computing along with an enormous number of statistical

analysis techniques that are employed in R package programs. Some important R software packages available for metabolomics analysis are MetabR, LiliKoi, MetaboAnalystR, and MetaboDiff.

9.12 Current and emerging methodologies of metabolomics in agriculture

The plant metabolome is considerably unpredictable containing more than 200,000 metabolites. These little atoms are very essential to analyze plant development and progression similarly as their response to natural changes. The flexibility of metabolomics in plant research is incredibly essential within the disclosure of biomarkers and in the improvement of harvest yield and quality. Unambiguous traceability of a crop ensures its origin, quality, and security. Untargeted metabolomics approaches have been extremely useful in the search for quality traits within crops. Ellis et al. (2018) performed untargeted metabolic profiling from genetically marked lines of *Pisum sativum* (pea) mature seeds. The study concerned the identification of genotypes that contained a high quantity of specific compounds associated with quality traits. The information contained relating to the sets of compounds in mature seeds associated with their genetic variation can be used to assist future breeding programs. The quality of seeds can be compromised within the crop by the occurrence of pests and diseases. The pea aphid *Acyrtosiphon pisum* comprises assorted biotypes that can affect *P. sativum* plants, each of them specialized on a particular crop legume species, including *Trifolium pratense* and *Medicago sativa*. The entire host races of this insect can develop themselves on *Vicia faba* (faba bean). For the identification of the metabolites involved in the specificity of pea aphid interaction with different host plants, Sanchez-Arcos et al. (2019) carried out a study on these four plant-herbivorous insect systems by making use of an untargeted metabolomics strategy. Focusing on the screening of phenolic compounds, a targeted strategy was chosen from pigmented and nonpigmented maize cultivars. High resistance to Fusarium infection was shown by the maize cultivar with the highest phenolic content, which became a promising result toward the selection of more resilient maize plants. Considering another plant-pathogen system, Chitarrini et al.'s work likewise demonstrated an incredible potential for a future application for the development of resistant varieties. Utilizing distinctive analytical techniques, they recognized biomarkers in a resistant grapevine associated with the defense against the biotrophic oomycete *Plasmopara viticola*, which is known to be the causative agent of downy mildew. This study helped to understand the mechanisms of grapevine interaction and resistance to downy mildew in a better way. A further step was taken by Negrel et al. in elucidating this grapevine *P. viticola* interaction and characterizing *P. viticola*'s metabolome by using an untargeted metabolomics approach. The utilization of these microbe biomarkers was done in the development of a monitoring assay for the early detection of *P. viticola* in grapevine. López-Gresa et al. characterized the profile of volatile organic compounds associated with the tomato immune response to these bacteria while analyzing both compatible and incompatible interactions between tomato (*Solanum lycopersicum*) and *Pseudomonas syringae*. These outcomes can be utilized later on in the development of safe tomato plants, hence preventing this agricultural problem and adding to more sustainable production. The availability of an optimal light environment is yet another concern for developing a stable fruit as this has been a major issue in several countries with lesser daylight hours. For the plants grown in greenhouses such as tomato, a supplementary light system is frequently utilized (Kaiser et al., 2019). To mark the metabolic changes in early fruit development of single-leaf tomato plants, Fukushima et al. performed an integrative omics approach with only one fruit truss, exposed to different intensities of red LED (light-emitting diode) light. The compounds that mostly contributed to the increase of fruit size of tomato plants and responded to the LED treatment were the metabolites that were mainly involved in the biosynthesis of several amino acids and carbohydrate metabolism. This is very pertinent given the significance of carbon allocation for fruits during their development, for which a reasonable source–sink relationship is essential to ensure sufficient fruit nutritional quality and yield. Beshir et al. investigate carbon reallocation changes throughout the development of apple fruit by utilizing isotopically labeled substrates and metabolomics (Beshir et al., 2017). Unexpectedly it was possible to make an intensive understanding of the metabolic dynamics occurring during the diverse developmental phases of fruit development utilizing dynamic isotope labeling experiments. The effective utilization of resources and productivity increase in a crop has been accomplished through controlled growth in plant industrial facilities. Even though these closed production systems are not a natural environment, yet they became more sustainable and attractive to the food industry with controlled lightning strategies and reduced environmental pollutants. To investigate how cultivation conditions affected leaf metabolic composition, Tamura et al. analyzed the metabolite profiles of lettuce leaves that were grown under hydroponic conditions or fertilized soil. It was seen in the results that the metabolic profile of both lettuce cultivars analyzed was enormously affected by the cultivation method. The affected metabolites are the ones responsible for taste and functional ingredients such as amino acids and phenolic compounds. Neugart et al. also analyzed the impact of soil fertilization with biological waste compost in the metabolic composition of *Brassica rapa* ssp. *Chinensis* (pak choi) sprouts. The addition of the biological waste from food production namely

coffee, hop, and aronia had highly affected sprout metabolic profile by decreasing the glucosinolates and phenolic compounds increasing the concentration of carotenoids. A detailed assessment of the effect of alternative cultivation systems, like plant factories and greenhouses from the studies of Tamura et al., Fukushima et al., and Neugart et al. shows that it is essential in the quality and nutritional value of crop products, especially the evaluation of the light, soil, and fertilization conditions. The techniques often selected for quantitative metabolomics analysis are NMR and LC–MS or GC–MS. However, when it comes to untargeted metabolomics, currently, Fourier-transform ion cyclotron-resonance MS (FT–ICR–MS) is used as it can simultaneously detect and identify a huge number of metabolites in high-throughput assays, providing high resolution as well as high accuracy in crop metabolomics (Allwood, Parker, Beckmann, Draper, & Goodacre, 2011).

9.13 Integration of metabolomics tools with other omics tools

Metabolomics tools can be integrated with various other omics tools, such as genomics, proteomics, and transcriptomics to tackle abiotic/biotic stresses in plants. Metabolomics devices are utilized for metabolic profiling of biofluids and different cell tissues, which are associated with various cell processes depicting the entire physiological composition of a cell (Kraly, Holcomb, Guan, & Henry, 2009). When compared to any other living species, a diverse range of plant metabolites have various orders of size, solvency, precariousness, unpredictability, extremity, and flexibility. Due to an insufficient connection between the proteome and metabolome, it is considerably challenging to elucidate plant metabolites in metabolic profiling. At certain times, due to technical hurdles like lack of standardized protocols, incompatibility of instruments, and volatility of the desired metabolites, it becomes difficult to detect some metabolites during the whole metabolome analysis. A set of different technologies are required to provide the greatest amount of metabolite coverage because no single technique or tool can be utilized to analyze the entire metabolites present in a metabolome. Different metabolomics techniques include nondestructive NMR spectroscopy, MS, CE–MS, high-performance thin-layer chromatography, LC–MS, GC–MS, ultraperformance liquid chromatography (UPLC), direct infusion MS, high-resolution MS, and FI–ICR–MS. Among these methods, GC–MS, CE–MS, LC–MS, and NMR-based integrated approaches are extensively used in metabolomic analysis.

For a thorough investigation of metabolites, the NMR procedure can be utilized in numerous living organisms, including plants as NMR-based metabolic profiling is fast, efficient and expedient, for the screening and identification of similar biological samples (Garcia-Perez et al., 2020). It is selective, nondestructive, and exceptionally proficient at mapping metabolic pathways. Also, its high reproducibility makes it a useful tool in plant metabolomics research. NMR-based metabolic profiling can also productively screen plant responses under biotic/abiotic stresses at different developmental stages. NMR along with other integrated techniques has been applied to identify the structural units of unknown metabolites. Isotope-labeled NMR, micro-coil NMR, and one- and two-dimensional NMR are the recently developed advanced tools for plant metabolomics. NMR is the main instrument that can recognize the particular labeling of stable isotopes (Deborde et al., 2017). NMR is a rapid, noninvasive, highly quantitative, and unbiased approach that requires minor sample preparation and no requirement for a chromatography separation process. When compared with MS, NMR has a lower dynamic range, poor sensitivity, and less resolution, in limited coverage of primary and secondary metabolites in plant metabolomics research. However, the major limitations in NMR technology have been overcome by recent developments like superconducting magnets, cryogenic probes, multidimensional NMR techniques, and miniaturized radiofrequency coils. The MS technique gives the advantage of quick sample preparation and assessment in their natural state. The conventional tools for metabolite analysis are ultrahigh-performance LC (UPLC) and HPLC. However, analytical platforms for plant metabolome profiling have been efficiently enhanced through the integration of these tools with MS. A high rate of sensitivity for metabolic profiling is shown by GC–MS analytical technology and it offers exceptional detection, separation, and identification because of the utilization of an electronic impact ionization point of supply. This method can also be utilized to probe primary metabolites, like organic acids, peptides, amino acids, sugars, alkaloids, ketones, lipids, esters, and sugar phosphate. The advantages of GC–MS include its high sensitivity, precision and resolution, reduced running cost, and speedy metabolic profiling (D'Amelia, Dell'Aversana, Woodrow, Ciarmiello, & Carillo, 2018). However, GC–MS has a drawback, that is, it can only be used to identify thermally unstable and volatile compounds. LC–MS technique uses an ESI source to analyze metabolites having high molecular weight, which are polar and thermolabile. It is executed to a great extent for secondary metabolite profiling, including glucosinolates, vitamins, flavonoids, and carotenoids, however, can likewise be utilized for primary metabolites' detection. LC–MS has an extraordinary feature, that is, without derivatization it can permit direct probing of metabolites in any sample. Both targeted and nontargeted methods perform LC–MS-based metabolic profiling. In the targeted technique a set of metabolites are identified and

quantified whereas, in the nontargeted approach, various types of chemical compounds are identified, like lipids, amino acids, and their derivatives. LC and MS are extensively used in plant metabolomics, integrated approaches for analytical research due to their higher accuracy and sensitivity. To achieve high-resolution imaging in metabolomics that demonstrates the arrangement patterns of metabolites in plant cells and tissues advanced tools of MS have been applied, such as matrix-assisted laser desorption ionization (Enomoto, Sensu, Yumoto, Yokota, & Yamane, 2018). Metabolomics has emerged as a more versatile strategy than genomics and proteomics with ongoing coordinated utilization of MS. Indeed, metabolic profiling of various crops, such as rice, wheat, sorghum, maize, and soybean, showed prominent applications of metabolomics in plant biology.

9.14 Metabolomics under normal and stress conditions in plants

A massive reduction in the global annual crop yield is because of the biotic and abiotic stresses that adversely affect crop productivity. It serves as a key to understanding the systems biology of plants by providing assistance in analyzing various exogenous and endogenous plant metabolites under extreme climatic stresses (Liang et al., 2018). Any change in plants' growth conditions that adversely affect plant metabolism, development, and physiology can be described as abiotic stress. It acts as a major limiting factor in agriculture production. The basic aim of exploring metabolic variations under abiotic stresses is to recognize various metabolites that permit the restoration of plant homeostasis and standardize metabolic changes. Besides, it is additionally used to probe specific compounds liable for offering abiotic stress resistance in plants. Tools like NMR, LC–MS, and GC–MS are extensively used in metabolomics studies to elucidate abiotic stress tolerance in plants (Ma et al., 2018). Abiotic stresses severely affect all essential mechanisms in plants from germination to maturity. Plant photosynthesis as well as the synthesis of all primary metabolites, including amino acids, sugar alcohols, and sugars, are badly affected and hampered by abiotic stresses. The main abiotic stresses include drought, low- and high-temperature, salinity, waterlogging, heavy metal, and chilling.

9.14.1 Drought stress

Drought is a major constraint for agricultural production around the world. Plants adopt several physiological modifications depending on their exposure to mild or severe drought stress such as greater nutrient uptake by plant roots, reduction in vegetative growth, leaf abscission, leaf area reduction, stomatal closure, and a decrease in the rate of photosynthesis. There may be variation in the timing and severity of water deficit ranging from long drought to short periods without rain at all (Lilley & Fukai, 1994). Plants synthesize many ubiquitous polyamines in response to drought stress as a defense mechanism. One of the important adaptation mechanisms to water deficit in several plants is the osmotic adjustment, in which active accumulation of solutes in response to drought takes place that ultimately, results in reduced osmotic potential, and contributes to maintaining cell turgor.

9.14.2 Salinity stress

Due to both natural processes and agricultural practices, the increased salinization of arable land is expected to increase ion toxicity and disturbance of the ion uptake mechanism and have a drastic impact on soil fertility, resulting in a high percentage of land loss by the middle of the century (Shabala & Munns, 2012). It will also lead to osmotic imbalance and cause metabolic syndrome that results in stunted growth and the capture of several physiological activities. Imbalanced Na⁺ ion concentrations cause ion toxicity, which not only hampers nutrient and water uptake in high salinity conditions but also affects the economically important crop species because of their sensitivity toward high salt concentration in the soil. High salinity in plants engenders both hyperionic and hyperosmotic stresses. To cope with salinity stress conditions, many primary and secondary metabolites are synthesized by plants.

9.14.3 Waterlogging stress

Another sort of abiotic stress is waterlogging that impedes crop development and yield (Ahmed et al., 2013). Due to the limited supply of CO₂ and oxygen, waterlogging causes extreme injuries to plants and eventually hampers the photosynthesis cycle. A higher duration of waterlogging causes hypoxia that prevents CO₂ assimilation and directly affects the roots. Waterlogging stress has three adaptations that is morphological changes metabolic alteration and signal transduction.

9.14.4 Temperature stress

One of the most crucial environmental factors for the determination of plant growth and development is temperature. An optimum temperature is essential for plant growth. Depending on the temperature fluctuation, plants suffer severe damage and the developmental processes are ceased. The homeostasis and other physiological mechanisms are disturbed by high temperatures (Thomason et al., 2018). However, there is a certain highly complex inducible mechanism that can help to extend the temperature range of survival in some species. Many secondary metabolites under heat stress, such as arachidic acid, alanine, allantoin, rhamnose, and myoinositol, are synthesized by plants.

9.14.5 Metal-induced stress

Another abiotic stress, the metal stress has become a significant factor that influences crop yield by showing variations signs in their molecular, biochemical, and physiological mechanisms (Foy, Chaney, & White, 1978). The high concentration of trace elements like zinc, cobalt, chromium, nickel, copper, vanadium, and tungsten are lethal to plants. Some of the major pollutants that influence plant stress are lead, zinc, chromium, cadmium, and nickel. When metal stress is in high concentration, it can cause growth arrest as well as cell death plants. It is due to cellular oxidation, metabolic retardation, and enzyme inhibition. Also, iron (Fe), copper (Cu), and manganese (Mn) play a significant role in plants biological processes (Ducic & Polle, 2005).

The tendency of plants to accumulate various kinds of metabolites in response to biotic stresses that are specific to tissues and species and act as biomarkers has helped to regulate biotic stress resistance in various plant species. In response to metabolomics profiling that determines the significant changes in primary and secondary metabolites of plants due to any pathogen attack, plants embrace various techniques to trigger defensive pathways against such pathogen attack. But it becomes difficult to decode the entire metabolome of a plant species due to the presence of highly diversified metabolites. Thus the compounds that are identified from biotic stressed plants help in looking for novel defense compounds and then serve as significant important plant defensive state markers. The expansion in the number of metabolites has been regarded as sensitive metabolic biomarkers in diverse plant species.

9.15 Applications and future perspective of metabolomics in plant biotechnology and agriculture

A remarkable place has been attained by metabolomics in plant biotechnology. Its expansion has a prominent effect on plant biology research. It has tremendous applications in plant sciences, ranging from the link between genotype and phenotype in response to climatic stresses, analyzing the cells biological mechanism, evaluating transgenic varieties, elucidating biosynthetic pathways, carrying out chemotaxonomic analyses to investigating various stresses and characterize cultivars, yet plant metabolomics requires a broader exploration and appropriate knowledge on data mining, processing, annotation, assessment, and evaluation (Yang et al., 2019). Genetic breeding is another application of metabolomics which has significantly reduced the varied time required for high-throughput genome sequencing and reverse genetics through metabolomics-assisted breeding. Exploration of complex metabolic pathways that administer significant regulatory processes in plant metabolism has been attained by the combination of various omics approaches including genomics, proteomics, transcriptomics, and metabolomics. Identification of metabolic markers and prediction of the size and nature of biotic/abiotic stress may be incorporated as a future application of metabolomics (Wolfender, Rudaz, Choi, & Kim, 2013). Discovering wide applications in crop improvement programs to develop high yield of crops, creating climate-smart crop varieties and stress-tolerant germplasm will require metabolomics-assisted breeding approaches. Other metabolomics approaches such as using of modern genome editing toolkits such as CRISPR/Cas9 system and speed breeding are certain fascinating areas where metabolomics is prepared for risk assessment associated with genetically engineered crops and do wonders for crop improvement (Razzaq, Saleem, et al., 2019).

In the present scenario the food consumed by humans relies on specific crops that are vital because of their nutritional value. Moreover, the possibility of detecting particular compounds with some peculiar function via metabolomics is essential to the endurance of a species. Since there are no constraints in the study of multiple species, there are numerous ways of using metabolomics in agriculture. However, the result of metabolomics from agriculture will affect different regions of study such as nutrition, medicine, genetic improvement, and food quality control. In the end, the statistical analysis and bioinformatics resources that are used to interpret the results prove to be critical to consider in a wide audit of metabolomics. The advancement in plant metabolomics has permitted the precise selection of desirable

traits. The innovation of technology from analyzing a single metabolite to high-throughput assays generating imprints of various metabolites has paved the way for the discovery/ development of better models for metabolite networks along with the identification of robust biomarkers (Führer & Zamboni, 2015).

The discovery of biomarkers has been pertinent among the modern challenges of metabolomics, particularly in terms of disease as it is important to detect, monitor, and treat them. Since the number of metabolites in chemical compounds is huge and sometimes remains undetectable due to lesser concentration, it becomes difficult to identify and quantify them in a reliable way. To solve such problems, new analytical techniques are developed to increase the detection of a wide range of metabolites based on their structural characteristics to identify even the lowest concentration of compounds (Patti, Yanes, & Siuzdak, 2012). Metabolomics is a tool to improve and enhance our understanding of the biochemistry and metabolism of the organism. In the last few years, its diverse applications have made the interpretation and analysis of results much easier. The incorporation of a wide variety of crops in this type of study is fundamental to know their qualities by considering the most essential trait from them and for developing an application that will be beneficial to food, health, and industry. In near future, there will be an increase in the field of study agriculture concerning metabolomic aspects to ensure world food sovereignty. As of now, it is important to carry out activities focused on their preservation and rational exploitation since most of the crops and their diversity are at high risk. The scope of metabolomics implementing various analytical techniques has allowed us to make use of its applications for the plant species along with the advancement of programs based on distinctive chemical traits. The integration of metabolomics and the other omics tools has highly improved the ability of a plant breeder to develop agronomically superior plants (Chaudhary et al., 2015).

References

- Ahmed, F., Rafii, M. Y., Razi Ismail, M., Shukor Juraimi, A., Rahim, H. A., Asfalza, R., & Abdul Latif, M. (2013). Waterlogging tolerance of crops: Breeding, mechanism of tolerance, molecular approaches, and future prospects. *BioMed Research International* (2013).
- Allwood, J. W., Parker, D., Beckmann, M., Draper, J., & Goodacre, R. (2011). *Fourier transform ion cyclotron resonance mass spectrometry for plant metabolite profiling and metabolite identification. Plant Metabolomics* (pp. 157–176). Humana Press.
- Allwood, J. W., Vos, R. C. D., Moing, A., Deborde, C., Erban, A., Kopka, J., . . . Hall, R. D. (2011). *Plant metabolomics and its potential for systems biology research: Background concepts, technology, and methodology, Methods in enzymology* (vol. 500, pp. 299–336). Academic Press.
- Alseekh, S., Tohge, T., Wendenberg, R., Scossa, F., Omranian, N., Li, J., Kleessen, S., et al. (2015). Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. *The Plant Cell*, 27(3), 485–512.
- Alvarez, S., Marsh, E. L., Schroeder, S. G., & Schachtman, D. P. (2008). Metabolomic and proteomic changes in the xylem sap of maize under drought. *Plant, Cell & Environment*, 31(3), 325–340.
- Aoki-Kinoshita, K. F. (2006). Overview of KEGG applications to omics-related research. *Journal of Pesticide Science*, 31(3), 296–299.
- Asakura, T., Sakata, K., Yoshida, S., Date, Y., & Kikuchi, J. (2014). Noninvasive analysis of metabolic changes following nutrient input into diverse fish species, as investigated by metabolic and microbial profiling approaches. *PeerJ*, 2, e550.
- Atkinson, N. J., & Urwin, P. E. (2012). The interaction of plant biotic and abiotic stresses: From genes to the field. *Journal of Experimental Botany*, 63(10), 3523–3543.
- Beleggia, R., Rau, D., Laido, G., Platani, C., Nigro, F., Fragasso, M., Vita, P. D., et al. (2016). Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernels. *Molecular Biology and Evolution*, 33(7), 1740–1753.
- Benfey, P., Small, I., Altmann, T., Bouchez, D., Casal, J., Crosby, B., Furner, I., et al. *The Multinational Coordinated Arabidopsis thaliana Functional Genomics Project Annual Report 2006. 2007*.
- Beshir, W. F., Mbong, V., Hertog, M. L. A. T. M., Geeraerd, A. H., den Ende, W. V., & Nicolai, B. M. (2017). Dynamic labeling reveals temporal changes in carbon re-allocation within the central metabolism of developing apple fruit. *Frontiers in plant science*, 8, 1785.
- Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., et al. (2004). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*, 9(9), 418–425.
- Bundy, J. G., Davey, M. P., & Viant, M. R. (2009). Environmental metabolomics: A critical review and future perspectives. *Metabolomics: Official Journal of the Metabolomic Society*, 5(1), 3.
- Causon, T. J., & Hann, S. (2016). Review of sample preparation strategies for MS-based metabolomic studies in industrial biotechnology. *Analytica Chimica Acta*, 938, 18–32.
- Chaudhary, J., Patil, G. B., Sonah, H., Deshmukh, R. K., Vuong, T. D., Valliyodan, B., & Nguyen, H. T. (2015). Expanding omics resources for improvement of soybean seed composition traits. *Frontiers in Plant Science*, 6, 1021.
- Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics*, 46(7), 714–721.
- Choi, M.-Y., Choi, W., Park, J. H., Lim, J., & Kwon, S. W. (2010). Determination of coffee origins by integrated metabolomic approach of combining multiple analytical data. *Food Chemistry*, 121(4), 1260–1268.
- Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., . . . Xia, J. (2018). MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, 46(W1), W486–W494.

- D'Amelia, L., Dell'Aversana, E., Woodrow, P., Ciarmiello, L. F., & Carillo, P. (2018). *Metabolomics for crop improvement against salinity stress, Salinity Responses and Tolerance in Plants* (Volume 2, pp. 267–287). Cham: Springer.
- Datta, S. K., Datta, K., Parkhi, V., Rai, M., Baisakh, N., Sahoo, G., Rehana, S., et al. (2007). Golden rice: Introgression, breeding, and field evaluation. *Euphytica*, *154*(3), 271–278.
- Dawid, C., & Hille, K. (2018). Functional metabolomics—a useful tool to characterize stress-induced metabolome alterations opening new avenues towards tailoring food crop quality. *Agronomy*, *8*(8), 138.
- De Castro, M. D. L., & Priego-Capote, F. (2010). Soxhlet extraction: Past and present panacea. *Journal of Chromatography. A*, *1217*(16), 2383–2389.
- De Filippis, L. F. (2018). *Underutilised and neglected crops: Next generation sequencing approaches for crop improvement and better food security. Global Perspectives on Underutilized Crops* (pp. 287–380). Cham: Springer.
- De Livera, A. M., Dias, D. A., Souza, D. D., Rupasinghe, T., Pyke, J., Tull, D., ... Speed, T. P. (2012). Normalizing and integrating metabolomics data. *Analytical Chemistry*, *84*(24), 10768–10776.
- Deborde, C., Moing, A., Roch, L., Jacob, D., Rolin, D., & Giraudeau, P. (2017). Plant metabolism as studied by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *102*, 61–97.
- Dong, X., Gao, Y., Chen, W., Wang, W., Gong, L., Liu, X., & Luo, J. (2015). Spatiotemporal distribution of phenolamides and the genetics of natural variation of hydroxycinnamoyl spermidine in rice. *Molecular plant*, *8*(1), 111–121.
- Ducic, T., & Polle, A. (2005). Transport and detoxification of manganese and copper in plants. *Brazilian Journal of Plant Physiology*, *17*(1), 103–112.
- Dunn, W. B., Bailey, N. J. C., & Johnson, H. E. (2005). Measuring the metabolome: Current analytical technologies. *Analyst*, *130*(5), 606–625.
- Ellis, N., Hattori, C., Cheema, J., Donarski, J., Charlton, A., Dickinson, M., Venditti, G., et al. (2018). NMR metabolomics defining genetic variation in pea seed metabolites. *Frontiers in plant science*, *9*, 1022.
- Enomoto, H., Sensu, T., Yumoto, E., Yokota, T., & Yamane, H. (2018). Derivatization for detection of abscisic acid and 12-oxo-phytodienoic acid using matrix-assisted laser desorption/ionization imaging mass spectrometry. *Rapid Communications in Mass Spectrometry*, *32*(17), 1565–1572.
- Farag, M. A. (2014). Comparative mass spectrometry & nuclear magnetic resonance metabolomic approaches for nutraceuticals quality control analysis: A brief review. *Recent patents on biotechnology*, *8*(1), 17–24.
- Fiehn, O. (2002). *Metabolomics—The link between genotypes and phenotypes. Functional genomics* (pp. 155–171). Dordrecht: Springer.
- Field, K. J., & Lake, J. A. (2011). Environmental metabolomics links genotype to phenotype and predicts genotype abundance in wild plant populations. *Physiologia Plantarum*, *142*(4), 352–360.
- Foy, C. D., Chaney, R. L. T., & White, M. C. (1978). The physiology of metal toxicity in plants. *Annual Review of Plant Physiology*, *29*(1), 511–566.
- Fuhrer, T., & Zamboni, N. (2015). High-throughput discovery metabolomics. *Current Opinion in Biotechnology*, *31*, 73–78.
- Fukushima, A., Kusano, M., Redestig, H., Arita, M., & Saito, K. (2009). Integrated omics approaches in plant systems biology. *Current Opinion in Chemical Biology*, *13*(5-6), 532–538.
- Garcia-Perez, I., Poma, J. M., Serrano-Contreras, J. I., Boulangé, C. L., Chan, Q., Frost, G., Stamler, J., et al. (2020). Identifying unknown metabolites using NMR-based metabolic profiling techniques. *Nature Protocols*, *15*(8), 2538–2567.
- Gong, L., Chen, W., Gao, Y., Liu, X., Zhang, H., Xu, C., ... Luo, J. (2013). Genetic analysis of the metabolome exemplified using a rice population. *Proceedings of the National Academy of Sciences*, *110*(50), 20320–20325.
- Gonzalez, N., Gévaudant, F., Hernould, M., Chevalier, C., & Mouras, A. (2007). The cell cycle-associated protein kinase WEE1 regulates cell size in relation to endoreduplication in developing tomato fruit. *The Plant Journal*, *51*(4), 642–655.
- Guijas, C., Rafael Montenegro-Burke, J., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., et al. (2018). METLIN: A technology platform for identifying knowns and unknowns. *Analytical Chemistry*, *90*(5), 3156–3164.
- Gupta, A., Naraniwal, M., & Kothari, V. (2012). Modern extraction methods for preparation of bioactive plant extracts. *International journal of applied and natural sciences*, *1*(1), 8–26.
- Hadacek, F. (2002). Secondary metabolites as plant traits: Current assessment and future perspectives. *Critical Reviews in Plant Sciences*, *21*(4), 273–322.
- Harbourne, N., Marete, E., Jacquier, J. C., & O'Riordan, D. (2009). Effect of drying methods on the phenolic constituents of meadowsweet (*Filipendula ulmaria*) and willow (*Salix alba*). *LWT-Food Science and Technology*, *42*(9), 1468–1473.
- Hegeman, A. D. (2010). Plant metabolomics—meeting the analytical challenges of comprehensive metabolite analysis. *Briefings in functional genomics*, *9*(2), 139–148.
- Hill, C. B., Taylor, J. D., Edwards, J., Mather, D., Langridge, P., Bacic, A., & Roessner, U. (2015). Detection of QTL for metabolic and agronomic traits in wheat with adjustments for variation at genetic loci that affect plant phenology. *Plant Science*, *233*, 143–154.
- Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., ... Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, N.Y.)*, *292*(5518), 929–934.
- Jackson, M. (1997). Hormones from roots as signals for the shoots of stressed plants. *Trends in Plant Science*, *2*(1), 22–28.
- Jamil, A., Riaz, S., Ashraf, M., & Foolad, M. R. (2011). Gene expression profiling of plants under salt stress. *Critical Reviews in Plant Sciences*, *30*(5), 435–458.
- Jansen, R. C., Tesson, B. M., Fu, J., Yang, Y., & McIntyre, L. M. (2009). Defining gene and QTL networks. *Current Opinion in Plant Biology*, *12*(2), 241–246.
- Kaddurah-Daouk, R., Kristal, B. S., & Weinshilboum, R. M. (2008). Metabolomics: A global biochemical approach to drug response and disease. *Annual Review of Pharmacology and Toxicology*, *48*(1), 653–683.

- Kaiser, E., Ouzounis, T., Giday, H., Schipper, R., Heuvelink, E., & Marcelis, L. F. M. (2019). Adding blue to red supplemental light increases biomass and yield of greenhouse-grown tomatoes, but only to an optimum. *Frontiers in plant science*, 9, 2002.
- Kim, H. K., & Verpoorte, R. (2010). Sample preparation for plant metabolomics. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques*, 21(1), 4–13.
- Kim, S., Ouyang, M., Jeong, J., Shen, C., & Zhang, X. (2014). A new method of peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data. *The annals of applied statistics*, 8(2), 1209.
- Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y., & Stitt, M. (2004). Metabolite profiling in plant biology: Platforms and destinations. *Genome Biology*, 5(6), 109.
- Kraly, J. R., Holcomb, R. E., Guan, Q., & Henry, C. S. (2009). Microfluidic applications in metabolomics and metabolic profiling. *Analytica Chimica Acta*, 653(1), 23–35.
- Li, Q., Zhao, C., Li, Y., Chang, Y., Wu, Z., Pang, T., ... Xu, G. (2011). Liquid chromatography/mass spectrometry-based metabolic profiling to elucidate chemical differences of tobacco leaves between Zimbabwe and China. *Journal of Separation Science*, 34(2), 119–126.
- Li, S., Tian, Y., Wu, K., Ye, Y., Yu, J., Zhang, J., Liu, Q., et al. (2018). Modulating plant growth–metabolism coordination for sustainable agriculture. *Nature*, 560(7720), 595–600.
- Liang, B., Ma, C., Zhang, Z., Wei, Z., Gao, T., Zhao, Q., ... Li, C. (2018). Long-term exogenous application of melatonin improves nutrient uptake fluxes in apple plants under moderate drought stress. *Environmental and Experimental Botany*, 155, 650–661.
- Liland, K. H. (2011). Multivariate methods in metabolomics—from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry*, 30(6), 827–841.
- Lilley, J. M., & Fukai, S. (1994). Effect of timing and severity of water deficit on four diverse rice cultivars III. Phenological development, crop growth and grain yield. *Field Crops Research*, 37(3), 225–234.
- Lopes, A. S., Santa Cruz, E. C., Sussulini, A., & Klassen, A. (2017). *Metabolomic strategies involving mass spectrometry combined with liquid and gas chromatography. Metabolomics: From Fundamentals to Clinical Applications* (pp. 77–98). Cham: Springer.
- Luo, J. (2015). Metabolite-based genome-wide association studies in plants. *Current Opinion in Plant Biology*, 24, 31–38.
- Ma, N. L., Lah, W. A. C., Kadir, N. A., Mustaqim, M., Rahmat, Z., Ahmad, A., ... Ismail, M. R. (2018). Susceptibility and tolerance of rice crop to salt threat: Physiological and metabolic inspections. *PLoS One*, 13(2), e0192732.
- Matsuda, F., Okazaki, Y., Oikawa, A., Kusano, M., Nakabayashi, R., Kikuchi, J., ... Saito, K. (2012). Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. *The Plant Journal*, 70(4), 624–636.
- Monteiro, M. S., Carvalho, M., Bastos, M. L., & Guedes de Pinho, P. (2013). Metabolomics analysis for biomarker discovery: Advances and challenges. *Current Medicinal Chemistry*, 20(2), 257–271.
- Montenegro-Burke, J. R., Aisporna, A. E., Paul Benton, H., Rinehart, D., Fang, M., Huan, T., Warth, B., et al. (2017). Data streaming for metabolomics: Accelerating data processing and analysis from days to minutes. *Analytical Chemistry*, 89(2), 1254–1259.
- Neuweger, H., Albaum, S. P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., ... Goesmann, A. (2008). MeltDB: A software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics (Oxford, England)*, 24(23), 2726–2732.
- Oldiges, M., Lütz, S., Pflug, S., Schroer, K., Stein, N., & Wiendahl, C. (2007). Metabolomics: Current state and evolving methodologies and tools. *Applied Microbiology and Biotechnology*, 76(3), 495–511.
- Pandey, A., Venkata Mohan, S., Chang, J.-S., Hallenbeck, P. C., & Larroche, C. (Eds.), (2019). *Biomass, Biofuels, Biochemicals: Biohydrogen*. Elsevier.
- Parry, M. A. J., & Hawkesford, M. J. (2012). An Integrated Approach to Crop Genetic Improvement F. *Journal of integrative plant biology*, 54(4), 250–259.
- Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Metabolomics: the apogee of the omics trilogy. *Nature Reviews. Molecular Cell Biology*, 13(4), 263–269.
- Piasecka, A., Kachlicki, P., & Stobiecki, M. (2019). Analytical methods for detection of plant metabolomes changes in response to biotic and abiotic stresses. *International journal of molecular sciences*, 20(2), 379.
- Pourmortazavi, S. M., & Hajimirsadeghi, S. S. (2007). Supercritical fluid extraction in plant essential and volatile oil analysis. *Journal of Chromatography. A*, 1163(1-2), 2–24.
- Razzaq, A., Sadia, B., Raza, A., Hameed, M. K., & Saleem, F. (2019). Metabolomics: A way forward for crop improvement. *Metabolites*, 9(12), 303.
- Razzaq, A., Saleem, F., Kanwal, M., Mustafa, G., Yousaf, S., Arshad, H. M. I., ... Joyia, F. A. (2019). Modern trends in plant genome editing: An inclusive review of the CRISPR/Cas9 toolbox. *International Journal of Molecular Sciences*, 20(16), 4045.
- Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. (2015). Computational and statistical analysis of metabolomics data. *Metabolomics: Official Journal of the Metabolomic Society*, 11(6), 1492–1513.
- Robinson, M. F., Heath, J., & Mansfield, T. A. (1998). Disturbances in stomatal behaviour caused by air pollutants. *Journal of Experimental Botany*, 461–469.
- Sanchez-Arcos, C., Kai, M., Svatoš, A., Gershenzon, J., & Kunert, G. (2019). Untargeted metabolomics approach reveals differences in host plant chemistry before and after infestation with different pea aphid host races. *Frontiers in Plant Science*, 10, 188.
- Sanderson, M. A., Skinner, R. H., Barker, D. J., Edwards, G. R., Tracy, B. F., & Wedin, D. A. (2004). Plant species diversity and management of temperate forage and grazing land ecosystems. *Crop Science*, 44(4), 1132–1144.
- Sato, S., Soga, T., Nishioka, T., & Tomita, M. (2004). Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *The Plant Journal*, 40(1), 151–163.
- Schauer, N., & Fernie, A. R. (2006). Plant metabolomics: Towards biological function and mechanism. *Trends in Plant Science*, 11(10), 508–516.

- Serkova, N. J., & Niemann, C. U. (2006). Pattern recognition and biomarker validation using quantitative ¹H-NMR-based metabolomics. *Expert Review of Molecular Diagnostics*, 6(5), 717–731.
- Shabala, S., & Munns, R. (2012). Salinity stress: Physiological constraints and adaptive mechanisms. *Plant stress physiology*, 1(1), 59–93.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
- Song, Y., Schreier, P. J., Ramírez, D., & Hasija, T. (2016). Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing*, 128, 449–458.
- Sugimoto, M., Kawakami, M., Robert, M., Soga, T., & Tomita, M. (2012). Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Current bioinformatics*, 7(1), 96–108.
- Sun, X., & Weckwerth, W. (2012). COVAIN: A toolbox for uni-and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics: Official Journal of the Metabolomic Society*, 8(1), 81–93.
- Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS Online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, 84(11), 5035–5039.
- Thomason, K., Ali Babar, M., Erickson, J. E., Mulvaney, M., Beecher, C., & MacDonald, G. (2018). Comparative physiological and metabolomics analysis of wheat (*Triticum aestivum* L.) following post-anthesis heat stress. *PLoS One*, 13(6), e0197919.
- Vaclavik, L., Lacina, O., Hajslova, J., & Zweigenbaum, J. (2011). The use of high performance liquid chromatography–quadrupole time-of-flight mass spectrometry coupled to advanced data mining and chemometric tools for discrimination and classification of red wines according to their variety. *Analytica Chimica Acta*, 685(1), 45–51.
- Wang, C., He, L., Li, D.-W., Bruschiweiler-Li, L., Marshall, A. G., & Brüschweiler, R. (2017). Accurate identification of unknown and known metabolic mixture components by combining 3D NMR with fourier transform ion cyclotron resonance tandem mass spectrometry. *Journal of Proteome Research*, 16(10), 3774–3786.
- Weckwerth, W., & Morgenthal, K. (2005). Metabolomics: From pattern recognition to biological interpretation. *Drug Discovery Today*, 10(22), 1551–1558.
- Wei, J., Wang, A., Li, R., Qu, H., & Jia, Z. (2018). Metabolome-wide association studies for agronomic traits of rice. *Heredity*, 120(4), 342–355.
- William Allwood, J., Ellis, D. I., Heald, J. K., Goodacre, R., & Mur, L. A. J. (2006). Metabolomic approaches reveal that phosphatidic and phosphatidyl glycerol phospholipids are major discriminatory non-polar metabolites in responses by *Brachypodium distachyon* to challenge by *Magnaporthe grisea*. *The Plant Journal*, 46(3), 351–368.
- Wolfender, J.-L., Rudaz, S., Choi, Y. H., & Kim, H. K. (2013). Plant metabolomics: From holistic data to relevant biomarkers. *Current Medicinal Chemistry*, 20(8), 1056–1090.
- Xiao, J. F., Zhou, B., & Ransom, H. W. (2012). Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *TrAC Trends in Analytical Chemistry*, 32, 1–14.
- Xu, Y., & Goodacre, R. (2012). Multiblock principal component analysis: An efficient tool for analyzing metabolomics data which contain two influential factors. *Metabolomics: Official Journal of the Metabolomic Society*, 8(1), 37–51.
- Yang, Q., Zhang, A.-h., Miao, J.-h., Sun, H., Han, Y., Yan, G.-l., . . . Wang, X.-j. (2019). Metabolomics biotechnology, applications, and future trends: A systematic review. *RSC Advances*, 9(64), 37245–37257.
- Zaynab, M., Fatima, M., Sharif, Y., Zafar, M. H., Ali, H., & Khan, K. A. (2019). Role of primary metabolites in plant defense against pathogens. *Microbial Pathogenesis*, 137, 103728.
- Zhang, A., Sun, H., Wang, P., Han, Y., & Wang, X. (2012). Modern analytical techniques in metabolomics analysis. *Analyst*, 137(2), 293–300.

Explore the RNA-sequencing and the next-generation sequencing in crops responding to abiotic stress

Éderson Akio Kido¹, José Ribamar Costa Ferreira-Neto¹, Eliseu Binneck², Manassés da Silva¹, Wilson da Silva, Júnior¹ and Ana Maria Benko-Iseppon¹

¹Federal University of Pernambuco, University in Recife, Brazil, ²Brazilian Agricultural Research Corporation, University in Recife, Brazil

10.1 Introduction

Transcriptomics is a relevant approach depicting the snapshot of the global gene expression profile of cell/tissue/organ or organism under a specific circumstance. Through transcriptomics, the RNAs, especially mRNAs (messenger RNAs) and regulatory RNAs, including noncoding RNAs (ncRNAs), such as small interfering RNAs and microRNAs, reveal how organisms modulate gene expression during a developmental situation or in responses to environmental stimulus or stress. In the early days of transcriptomics and first-generation sequencing (the Sanger-sequencing method), a limited number of partial transcripts (e.g., ESTs—expressed sequence tags) already indicated the potential of this approach. Since then, several techniques (with open or closed architecture) have been developed and applied. However, many studies lacked depth and involved experimental designs that did not provide robust statistical support to assess the global gene expression. Despite many possibilities, the RNA-Seq (sequencing of RNA) method associated with next-generation sequencing (NGS) techniques (sometimes also called second-generation sequencing, SGS) is undoubtedly the most accurate and disseminated transcriptomic approach applied to crops nowadays. Here we present an RNA-Seq overview based on scientific articles covering plant abiotic stress responses published in the last 5 years, together with a commented overall RNA-Seq analysis workflow. We hope this chapter provides information helping breeders make good decisions favoring the improvement of crops worldwide.

10.2 From the beginning to the crop sciences: transcriptome analysis, its evolution, and state of the art

From the first RNA-Seq (77-nt yeast alanine tRNA) (Holley et al., 1965) until the genome-wide transcriptome analysis, almost half a century of improvements and technological advances have passed. Initial transcriptome sequencing efforts focused on viruses (Fiers et al., 1976) and simple eukaryotes, such as yeasts (Holley et al., 1965), due to the higher genetic complexity of cultivated plants. So, in the 1970s, Fiers et al. (1976) sequenced the complete transcriptome (3569 nucleotides) of the Ms2 bacteriophage, pioneering the transcriptome sequencing era (Fig. 10.1).

At that moment, similar studies with complex organisms were not feasible due to the unstable nature of RNAs, revealing technical limitations. However, after the discovery of the reverse transcriptase enzyme (Temin & Mizutani, 1970), synthesizing from RNAs the respective complementary DNAs (cDNAs), which are molecules with higher structural stability, the reverse transcription (RT) allowed access to the whole transcriptome.

Taking advantage of the RT protocol, Adams et al. (1991) initiated a systematic human cDNA sequencing project (Fig. 10.1), applying the Sanger-sequencing method on ABI 373A automatic DNA sequencers (Applied Biosystems, Inc.). This pioneering use of the first-generation DNA sequencing technology provided a batch of sequences (around 400 bp), named “ESTs,” functioning as substrates for contigs and transcriptome mapping.

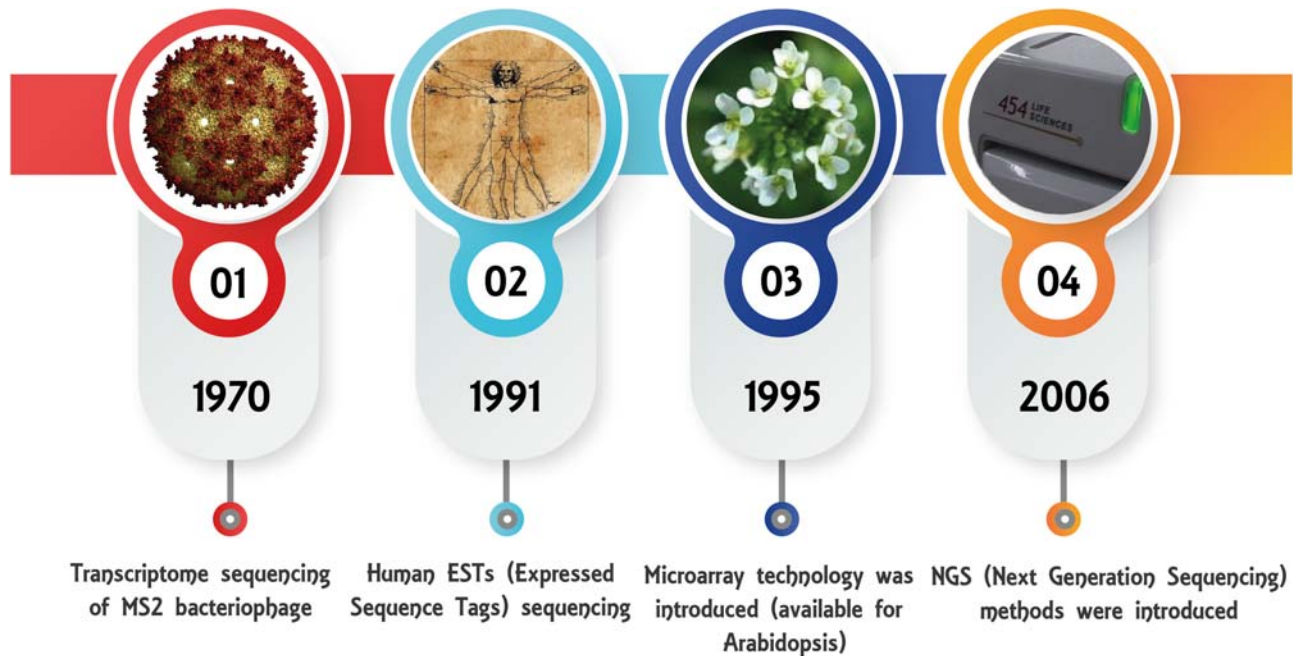


FIGURE 10.1 The timeline with four cornerstone events associated with transcriptomics studies.

Another technical innovation in transcriptome analysis was the microarray technology (Schena, Shalon, Davis, & Brown, 1995) (Fig. 10.1). The method, developed by Mark Schena and collaborators, consisted of monitoring hybridization events on a flat surface with immobilized DNA sequences encoding specific transcripts (the probes). The detection of target-probe hybridizations allows determining the relative abundance of nucleic acid sequences related to the targets. However, due to its closed architecture, the microarray technology needs the reference genome and the transcriptome before designing the microarray (chip). In the original publication, the authors described the robotic printing of DNA elements (representing 45 *Arabidopsis* genes) to the surface of a silane-coated glass microscope slide (Schena et al., 1995).

Although the mentioned techniques (EST and microarray) represented relevant milestones of scientific progress in gene expression analysis, some limitations concerning throughput, chemistry, time, costs, and accessibility hampered a more robust and complete genome-wide transcriptome analysis. For details, see the review of Murphy (2002) and also Lorkowski and Cullen (2006).

With the development of sequencing methods called “NGS” (Fig. 10.1), some of those limitations have been overcome, and next-generation sequencers achieving around a 100-fold increase in throughput over the Sanger-sequencing method showed at that time (around 2005) substantial improvements in quality and yield (Margulies et al., 2005). Thus the NGS technology proved to be a valuable tool in biological research, with tremendous potential for global gene expression profiling.

Initially, several molecular approaches took advantage of the high-throughput (HT) capacity of the NGS technology. For instance, the original transcriptomic SAGE (serial analysis of gene expression) method (Serial Analysis of Gene Expression; Velculescu, Zhang, Vogelstein, & Kinzler, 1995), providing gene expression profiles based on tags (11–14 bp) extracted from cDNAs, basically employed the Sanger-sequencing approach. Similar procedures also involved the improved LongSAGE (tags of 20 bp; Saha et al., 2002) and SuperSAGE (with tags of 26 bp; Matsumura et al., 2005), both techniques presenting better tag length, which provided better tag-annotation.

The SuperSAGE method combined with the NGS approach resulted in the HT-SuperSAGE technique (Matsumura et al., 2010), allowing a deepSuperSAGE analysis (Molina et al., 2011), which promotes the genome-wide transcriptome analysis. Kido, Ferreira Neto, Kido, Pandolfi, and Benko-Iseppon (2013), applying a user-friendly bioinformatics approach with DeepSuperSAGE data, contrasted two cowpea accessions [*Vigna unguiculata* (L.) Walp.] showing different drought-tolerance phenotypes and identified candidate genes responding differentially to abiotic stress of roots dehydrated after 150 minutes. Another tag-based NGS method, named Massive Analysis of cDNA Ends (Zawada et al., 2014), also provided a “digital gene expression profiling” based on tags, but differentially of the SAGE or its improved methods, without using any tagging enzyme.

However, nowadays, due to the costs of nucleic acid sequencing becoming cheaper and faster, the tag-based NGS approach has become less employed than the RNA-Seq approach. The RNA-Seq technique emerged just over a decade ago as another NGS application, and has become one of the most ubiquitous tools in molecular biology today, revolutionizing biological research in the 21st century (Nagalakshmi et al., 2008).

Marioni, Mason, Mane, Stephens, and Gilad (2008) proposed an RNA-Seq protocol with statistical gene expression analysis, the results of which were compared with those from gene expression arrays, the considered gold standard test. Identifying new transcripts and studying gene expression profiling in a wide-transcriptome approach remain the principal uses of the RNA-Seq method. Thus to know how plant RNA-Seq studies have been performed in the last 5 years (since 2015), we conducted a data mining in the PubMed database (<http://pubmed.ncbi.nlm.nih.gov>; September 2020), the results of which are commented next.

10.3 The overview on plant sequencing of RNA studies

The data-mining strategy with the keywords “RNA-Seq AND plants AND abiotic stress” performed in the PubMed database highlighted 1102 scientific articles published in the last 5 years with plant-model species and crops. About crops, the data mining identified 389 scientific reports. From these, 176 also reported the generation of the analyzed RNA-Seq libraries, and these articles composed the basis for the RNA-Seq overview presented in Fig. 10.2.

The identified manuscripts compiled concerning 13 issues disclosed next provided the RNA-Seq overview, presented in Fig. 10.2, and discussed here:

1. crops: the most studied crops identified in the publications were maize (9.7%, e.g., Zenda et al., 2019), rice (9.1%, e.g., Chung et al., 2016), and *Brassica napus* (5.1%, e.g., Ma et al., 2017). Maize and rice are two food crops of world relevance; together with wheat, they provide at least 30% of the food calories to more than 4.5 billion people in 94 developing countries (Shiferaw, Prasanna, Hellin, & Bänziger, 2011);
2. plant tissues/organs: the identified RNA-Seq libraries comprised basically roots (e.g., Zhao et al., 2018) and leaves (e.g., Peng et al., 2014). Roots are plant organs with great adaptive capacity, being able to grow and carry out their development under different environmental conditions (e.g., substrates, humidity); usually, roots are molecularly characterized in plants responding to abiotic stresses (for a review, see Sánchez-Romera & Aroca, 2020). Leaves are plant organs where photosynthesis takes place (in the mesophyll); according to Chaves, Flexas, and Pinheiro (2009), salt- and drought-stress effects on plant photosynthesis are direct (e.g., limiting CO₂ diffusion through stomata and the mesophyll, altering the photosynthetic metabolism) or indirect (e.g., the oxidative stress from multiple stresses imposition); the carbon balance of a plant during a period of salt/drought stress and recovery may depend on the velocity and degree of photosynthetic recovery (Chaves et al., 2009);
3. abiotic stress: most plants analyzed in Seq-RNA assays were subjected to drought (25%), salinity (20%), or cold (11%); manuscripts covering drought and salinity reflect global warming on climate change trends (Swann, 2018) and the increase in soil salinization processes worldwide (Shahid, Zaman, & Heng, 2018); about drought, the imposed experimental method varied, and also the methods of drought-stress characterization; some articles reported polyethylene glycol (e.g., Moon et al., 2018), an osmotic-stressing agent to plant cells, promoting plant dehydration, while others indicated visual drought symptoms, characterizing the phenotypic manifestation of the applied water-deficit treatment (e.g., Danilevskaya et al. (2019) suppressed the irrigation of corn plants, collecting samples when 50% of the treated plants showed leaf wilting); another strategy defines the exposition time of plants to the stress in question (e.g., Zhang et al., 2014); some strategies observed in experimental assays simulating drought in plant transcriptomic studies were reviewed by Kido, Ferreira-Neto, Pandolfi, de Melo Souza, and Benko-Iseppon (2016). Concerning salinity, the NaCl molarities varied from 50 μ M (e.g., Zhang et al., 2018) to 400 μ M (e.g., Wu et al., 2020); regarding cold, plants were exposed to 10°C (e.g., Xu, Zhang, Liu, Yang, & Hou, 2016), and even to -20° C (e.g., Mousavi et al., 2014);
4. abiotic stress exposure time: the duration and timing of the stress period imposed on plants analyzed by RNA-Seq approaches varied according to the nature and the level of the applied stress, but basically covered days (43%, e.g., Morgil, Tardu, Cevahir, & Kavakli, 2019), days and hours (25%, e.g., Wang et al., 2016), or only hours (18%, e.g., Ma et al., 2019); however, two publications reported 30 minutes. as the exposure time of treated plants (Dang et al., 2013; Wan et al., 2015);
5. gene expression profiling: most of the data-mined publications (93%) highlighted global transcriptome analyses (e.g., Wu et al., 2020), but in 7% of the articles, the authors analyzed specific categories or components expressed in the plant responses; for example, Tang et al. (2019) performed a genome-wide identification and expression

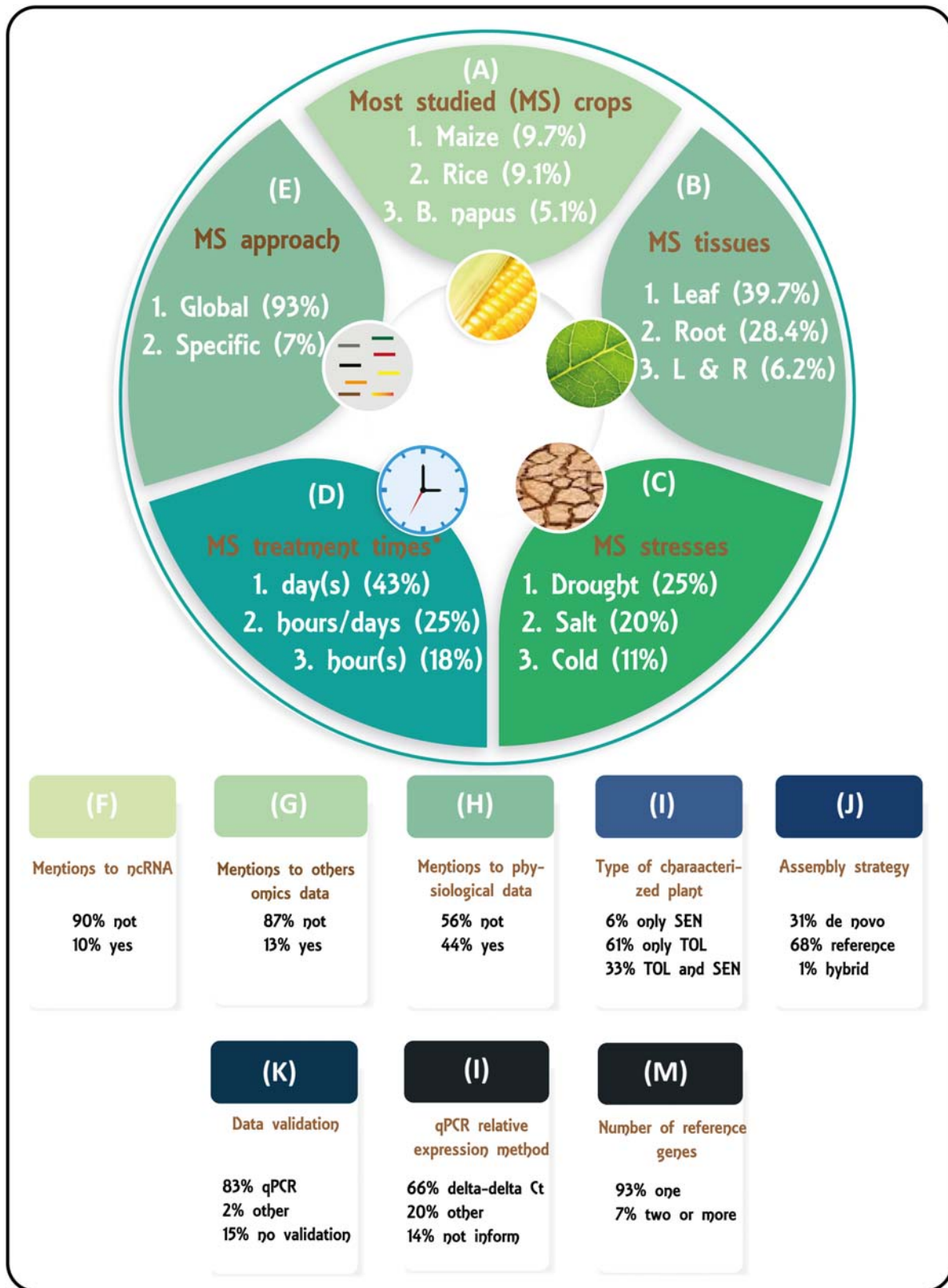


FIGURE 10.2 Identified manuscripts using the keywords “RNA-Seq AND plant crops AND abiotic stress” in the PubMed database (<https://pubmed.ncbi.nlm.nih.gov/>) and its distribution according to 13 issues. *Only manuscripts with details of the RNA-Seq library generation and published since 2015. *RNA-Seq*, Sequencing of RNA.

profile in physic nut plants, analyzing only HD-ZIP genes codifying transcription factors (TF) of that family, while [de Lima Cabral et al. \(2020\)](#) reported the first TFome (23 TF families), based on differentially expressed genes of physic nut responding to NaCl (150 mM);

6. ncRNAs: most of the identified plant RNA-Seq studies (90%) reported only transcripts encoding proteins (e.g., [Wu et al., 2015](#)); however, in the remained studies (around 10%), ncRNAs are also mentioned; [Di Bella et al. \(2019\)](#) reviewed some pipelines developed to detect specifically ncRNAs from RNA-Seq data;
7. other omics data: as expected, most identified manuscripts reported only RNA-Seq data (90%); however, in the remained articles also genomic data are presented, usually associated with genes differentially expressed (e.g., [Zhao et al., 2018](#));
8. physiological data: most of the identified RNA-Seq publications (56%) presented the desirable physiological characterization of stress-treated plants (e.g., [Cui et al., 2019](#)), but an expressive number of publications showed no similar characterization (e.g., [Wu et al., 2019](#));
9. studied accession: in most of the identified RNA-Seq studies (67%), the studied phenotype corresponded to the abiotic stress-tolerant accession (e.g., [Hübner, Korol, & Schmid, 2015](#)); the tolerant gene expression profile was explored in such cases, looking for gene/transcript associated with the tolerant phenotype; however, in the remained studies (33%), the authors also included the stress-sensitive accession (e.g., [Huang et al., 2019](#)), maximizing the biological information generated;
10. transcriptome assembly: the strategy applied in most of the data-mined RNA-Seq manuscripts (68%) was the genome-guided assembly (e.g., [Wang et al., 2018](#)), highlighting for the respective crops the genome availability, whereas for the remaining publications, the strategy employed was de novo assembly; details about the two assembly strategies are presented in the RNA-Seq workflow discussed in the present chapter;
11. gene expression data validation: in most of the plant RNA-Seq studies (83%), authors employed the qPCR (quantitative real-time PCR) method (e.g., [Yang & Huang, 2018](#)); however, in 2% of the publications, the authors applied a different technique, but, unfortunately, in 15% of the manuscripts, gene expression data were not validated;
12. relative quantification method: in the qPCR assays, most of the relative quantification applied the $\Delta\Delta C_t$ method (66% of the manuscripts, e.g., [Zhao, Wei, Ji, & Ma, 2019](#)), while in 20% of those qPCR studies, the authors performed a different method, but an expressive amount of the manuscripts (14%) did not inform any validation process;
13. reference genes in the qPCR assays: in most of the identified manuscripts (93%), the authors employed a single reference gene (endogenous control) to normalize for variations of sample loading (e.g., [Wang et al., 2016](#)), while in only 7% of the related manuscripts, the authors employed two or more reference genes; regarding the mentioned issue and the previous one, most of the authors did not follow the MIQE guidelines (Minimum Information for Publication of Quantitative Real-Time PCR Experiments; [Bustin et al., 2009](#)); such guidelines target the reliability of results to help ensure the integrity of scientific literature, promote consistency between laboratories, and increase experimental transparency.

After presenting the RNA-Seq overview covering cultivated plants exposed to abiotic stresses, we provide some information about an overall RNA-Seq analysis workflow.

10.4 The RNA-sequencing analysis workflow

Here we describe the main steps of a typical differential gene expression study with RNA-Seq data, concerning the main focus on data analysis, as illustrated in [Fig. 10.3](#). In addition, a non-exhaustive list of programs used in different stages of RNA-Seq analysis is shown in [Table 10.1](#). The instructions for users concerning the mentioned tools are available in the software documentation, related references, or online pages.

10.4.1 Data generation

The foundation of a typical study covering differential gene expression and RNA-Seq is data generation. Initially, it is essential to clearly establish the research's biological questions and adequately define the experimental design. An excellent experimental design does not have to be complicated. Still, it is essential to note that no analysis, no matter how impressive, can remedy an experiment conducted with an inappropriate design. That is why it is crucial to involve the whole team from the beginning of the study ([Glass, 2014](#); [Quinn & Keough, 2002](#)). When defining the experimental

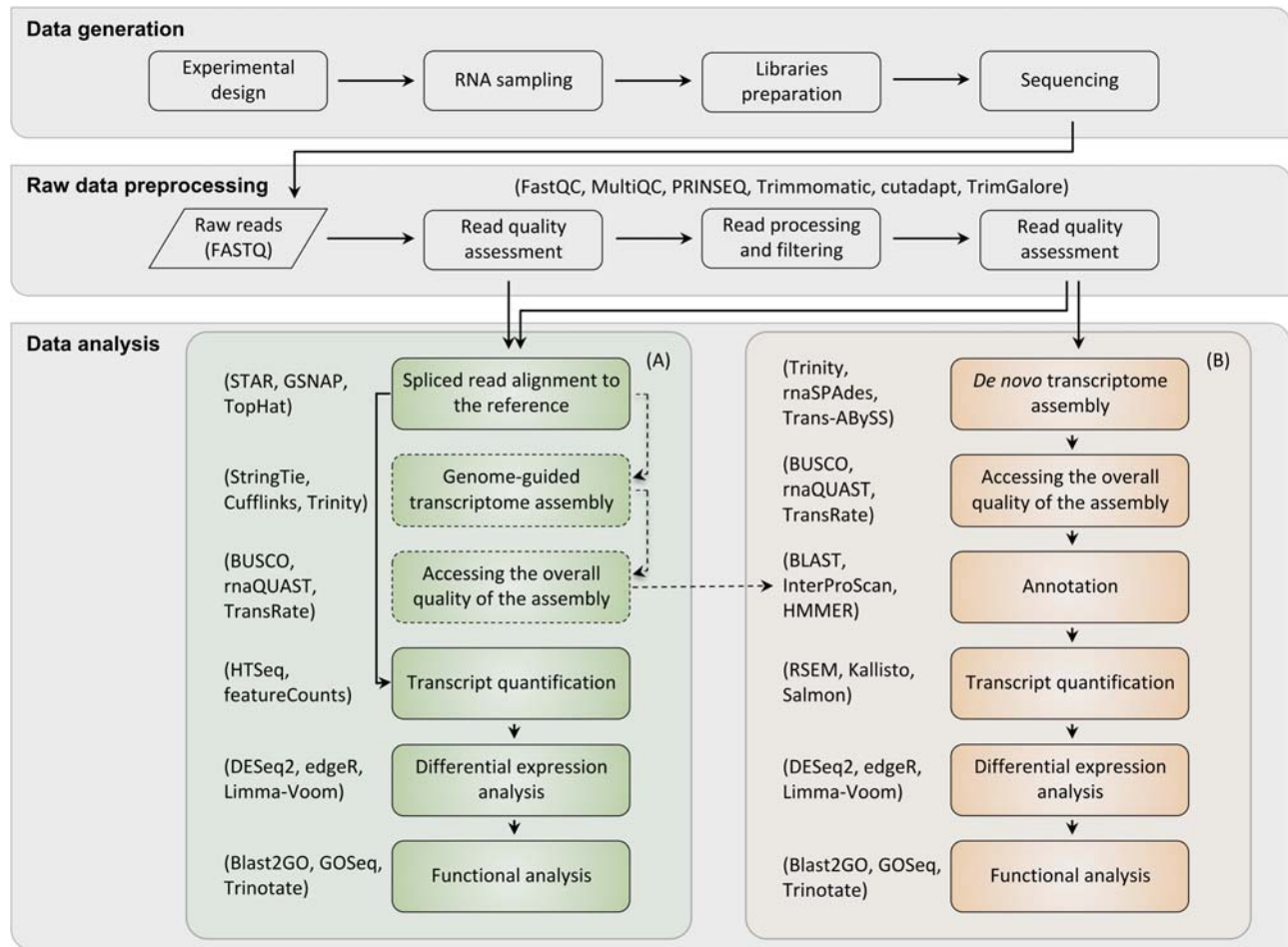


FIGURE 10.3 An overview of the paths to be followed in a typical gene expression analysis by RNA-Seq approach: the workflow left (A) is preferred for data analysis when a reference genome is available for the species, and the workflow right (B) when the reference genome is not provided. *RNA-Seq*, Sequencing of RNA.

design, the primary issues are the number of replicates, biological and technical, the sequencing depth (read depth), the desired read length, and the cDNA sequencing approach, providing single-end or paired-end reads.

As an experimental technique involving appropriate statistical analysis, RNA-Seq analysis requires replication to infer the variability between and among groups. It is widely accepted that the greater the number of replicates in an RNA-Seq assay, the more robust the results. Unfortunately, the number of replicates tends to be minimized due to financial limitations, including the cost of RNA extraction/preparation, library generation, and sequencing of libraries. However, an insufficient number of replicates can impair the data's quality, compromising the biological interpretation of the results.

Biological replicates serve the study better than technical replicates since the objective is to capture the natural variability between different biological samples within an experimental group. Technical replicates involving the same initial biological sample are generally unnecessary. But, they can be useful for assessing the technical reproducibility of the sequencing process.

In practice, it has traditionally been used as a rule of thumb to recommend a minimum of three biological replicates (Conesa et al., 2016). However, for the experiment to be quite reliable, it is best to estimate the minimum number of replicates required based on biological variability between samples, the technical variability in the sequencing procedures, and the intended statistical power. These values are generally not available *a priori* but can be derived from similar public datasets. From a dataset, one can estimate, for example, the power of their experimental design for a given method of differential expression analysis (Conesa et al., 2016; Lamarre et al., 2018; Wu & Wu, 2016), using R packages such as PROPER (Wu, Wang, & Wu, 2015) and RnaSeqSampleSize (Zhao, Li, Guo, Sheng, & Shyr, 2018).

TABLE 10.1 Software tools currently used in RNA-Seq (sequencing of RNA) analysis workflow to discover differential gene expression.

Tool name	Use category	Website (URL)
FastQC	Read quality assessment	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC	Read quality assessment	https://multiqc.info
PRINSEQ	Read quality assessment	http://prinseq.sourceforge.net
Trimmomatic	Read processing and filtering	http://www.usadellab.org/cms/index.php?page=trimmomatic
Cutadapt	Read processing and filtering	https://cutadapt.readthedocs.io/en/stable/
TrimGalore	Read processing and filtering	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
STAR	Spliced read alignment to the reference	https://github.com/alexdobin/STAR
GSNAP	Spliced read alignment to the reference	https://github.com/juliangehring/GMAP-GSNAP/blob/master/INSTALL
TopHat	Spliced read alignment to the reference	http://ccb.jhu.edu/software/tophat/index.shtml
StringTie	Genome-guided transcriptome assembly	https://ccb.jhu.edu/software/stringtie/
Cufflinks	Genome-guided transcriptome assembly	http://cole-trapnell-lab.github.io/cufflinks/
Trinity	De novo and genome-guided transcriptome assembly	https://github.com/trinityrnaseq/trinityrnaseq/wiki
rnaSPAdes	De novo transcriptome assembly	https://cab.spbu.ru/software/rnaspades/
Trans-ABYSS	De novo transcriptome assembly	https://www.bcgsc.ca/resources/software/trans-abyss
BUSCO	Assessing the overall quality of the assembly	https://busco.ezlab.org
rnaQUAST	Assessing the overall quality of the assembly	https://cab.spbu.ru/software/rnaquast/
TransRate	Assessing the overall quality of the assembly	https://hibberdlab.com/transrate/
BLAST	Annotation	https://www.ncbi.nlm.nih.gov/books/NBK279690/
InterProScan	Annotation	https://www.ebi.ac.uk/interpro/
HMMER	Annotation	http://hmmer.org/
HTSeq	Transcript quantification	https://htseq.readthedocs.io/en/release_0.9.1/index.html
featureCounts	Transcript quantification	http://bioinf.wehi.edu.au/featureCounts/
RSEM	Transcript quantification	http://deweylab.github.io/RSEM/
Kallisto	Transcript quantification	http://pachterlab.github.io/kallisto/
Salmon	Transcript quantification	https://salmon.readthedocs.io/en/latest/salmon.html
DESeq2	Differential expression analysis	http://bioconductor.org/packages/release/bioc/html/DESeq2.html
edgeR	Differential expression analysis	http://bioconductor.org/packages/release/bioc/html/edgeR.html
Limma-Voom	Differential expression analysis	http://bioconductor.org/packages/release/bioc/html/limma.html
Blast2GO	Functional analysis	https://www.blast2go.com
GOSeq	Functional analysis	https://bioconductor.org/packages/release/bioc/html/goseq.html
Trinotate	Functional analysis	https://github.com/Trinotate/Trinotate.github.io/wiki

Concerning the read depth (the number of sequencing reads obtained for a sample), sufficient sequencing coverage is relevant to detect the different transcripts expressed at different levels in each experimental condition. Therefore, as well as the number of biological replicates, the estimate of the read depth metric takes into account the research

approach and some biological characteristics of the organism, such as reference genome availability, genome size, ploidy level, and gene number.

It is equally important to consider whether the study aims to analyze differential expression by gene alone or by transcript and isoform (distinct transcripts resulting from alternative splicing events). Typically, each library (representing a sample) shows a read depth of 10–30 million reads. However, the ideal read count per sample depends on the species under study and the variables listed earlier.

If excessively increased, the read count tends to give progressively smaller returns, while the decrease may limit the power to detect differentially expressed genes (Bass, Robinson, & Storey, 2019; Liu, Zhou, & White, 2014). The bottom line implies that, given some replicates and a read depth, especially when the transcriptome will be de novo assembled, the minimal sequencing coverage depth required needs to be satisfied (Li, Tong, Xia, & Wei, 2019). However, it is better to increase the number of replicates than to massively increase the depth of read coverage of an insufficient number of replicates. That is because biological replicates contribute more to improving the statistical significance tests' robustness to detect relatively lower changes in gene expression at the transcript level (< twofold).

After sample collection, the total RNA is extracted and prepared for the library construction (mRNA enrichment, fragmentation, cDNA synthesis, adapter ligation, PCR amplification). The library sequencing is usually carried out by a certified sequencing service provider, which returns the raw sequencing data for subsequent processing and analysis.

Concerning the library construction protocols, the first-generation RNA-Seq protocol did not preserve that information about which strand originated the transcripts. In the strand-specific library preparation protocol or stranded RNA-Seq, the specificity about the origin for each transcript remains in the sequencing data (Levin et al., 2010). Although relatively more expensive, it is preferable to perform a stranded protocol because the strand information provides greater specificity in quantifying the transcript expression (Zhao et al., 2015), correct assembly of the transcripts, and better accuracy of new transcripts identification.

The cDNA sequencing is performed from just one end (single-end) or both ends (paired-end), and the choice between single-end or paired-end sequencing depends primarily on financial constraints since paired-end sequencing implies sequencing twice as many reads. Considering that specific size fragments were selected in library construction, paired-end sequencing provides more information for the alignment process, benefiting the read mapping to reference for the genome-guide or de novo assembly, especially the last one.

About the read length, 50 bases are satisfactory for differential gene expression analysis, which is the minimum recommended. Regarding de novo transcriptome assembly, longer read length usually benefits the results when good-quality reads are provided. Additionally, in paired-end sequencing, it is usual to determine the read length so that the total fragment size in the sequencing library is slightly higher than the sum of two read lengths.

Concerning the sequencing platforms, Stark, Grzelak, and Hadfield (2019) reviewed potentialities and applications of the main NGS technologies available for RNA-Seq, including short-read cDNA sequencing (Illumina and Ion Torrent platforms), long-read cDNA sequencing (Pacific Biosciences—PacBio and Oxford Nanopore—ONT), and long-read direct RNA sequencing (ONT platform). The authors inferred, and we also concluded, that more than 95% of the published RNA-Seq data available today applied the Illumina short-read sequencing technology. Thus Illumina is currently the technology of choice for RNA-Seq to detect and quantify transcriptome-wide gene expression, mainly due to the HT and relatively low-error rates. Therefore the RNA-Seq analysis described next concerns the Illumina platform.

10.4.2 Raw data processing

The sequencing process is susceptible to read errors, bias, artifacts, and adapter/primer contamination, requiring adequate preprocessing before submitting the data to the transcriptome assembly. Sequencing errors are generally characterized as low-quality sequences, meaning that the read sequence analyzed is not faithful to the original sequence. Typically, the end of the reads accumulates most of the sequencing errors, and these errors provide differences in reads regarding the original sequences, making it difficult to assemble the full-length transcripts. With a trimming process step, the last low-quality bases are pruned, preparing the reads for assembly.

Contaminants such as vectors, adapters, and polyadenylated segments at the end of the reads also interfere with the assembly, possibly joining sequences from different transcripts in the same contig. Even considering that most assemblers perform detection and filtering of errors and, sometimes, correction of contig error, such errors and contaminants must be detected and properly removed before assembly.

Therefore filtering low-quality reads and contaminants and trimming low-quality bases and adapter sequences are essential to speed up and optimize the further steps in the analysis, especially for de novo transcriptome assembly. It is

crucial to perform a read quality assessment before and after filtering and trimming reads to evaluate the results of the processing steps.

10.4.3 Data analysis

10.4.3.1 Step 1—transcriptome assembly

The transcriptome assembly aims to establish the set of transcripts depicting the transcriptome of the organism under study. Even in the case of model organisms and well-studied species with a fully sequenced and annotated genome, the transcriptome assembly may be necessary, for instance, to discover new transcripts and splicing variants.

It is common to classify assembly methods in two categories: (1) *de novo* when the assembly of transcripts includes only the reads and (2) genome-guided when the assembly reflects the spliced alignment of the reads with the reference genome. The approach to be chosen depends mainly on the availability of the reference genome. However, many crops lack a reference genome, and the RNA-Seq study depends on the *de novo* approach. In addition to the cases when the reference genome is not available, the *de novo* approach is performed complementing the genome-guided strategy or when the genome sequence is very fragmented.

The *de novo* assembly approach reconstructs the transcripts considering the overlapping of sequencing reads. If the bases at the end of a read match to the bases at the end(s) of another read(s), then connection and order between them are established about the original sequence. In most modern *de novo* assembly algorithms, these orders involving the sequencing reads, represented in a *de Bruijn* graph (Flicek & Birney, 2009; Geniza & Jaiswal, 2017), are analyzed for the assembly of contigs, representing the transcripts' sequences. The comparison of contigs reduces redundancies and allows artifact identification, such as transcript fusion and spurious insertions. Ideally, full-length transcripts are reconstructed, but this is hardly achieved, and most of the transcripts are usually partial sequences (fragments).

Relevant challenges in *de novo* assembly are the correct assembly of full-length transcripts and specific splicing isoforms, as those low expressed, and the distinction between closed members of gene families (paralogs and orthologs). Although the *de novo* approach is relatively more challenging to perform than the genome-guided strategy, requiring higher computational processing power, *de novo* assemblers are in continuous development, implementing new error checking routines to provide more accurate results (Hölzer & Marz, 2019).

When a high-quality reference genome sequence is available, the recommended genome-guided approach first aligns sequencing reads to the reference genome using a specialized algorithm for spliced alignment, identifying clusters of reads representing the potential transcripts. Thus, based on the alignment results, the assembly is generated.

Genome-guided approaches about *de novo* strategies usually generated better assemblies, showing a better distinction between paralogs, higher sensitivity detecting low-expressed splice variants, and lower artifact generation, such as transcript fusion and spurious insertions. However, to ensure these advantages, it is necessary to provide a high-quality reference genome sequence with minimal errors and well-annotated exons. The final result also depends on the accuracy of the alignments and mapping of reads in the provided reference genome.

As mentioned, the *de novo* strategy may complement the genome-guided approach, contributing to solving problems, such as covering gaps in the reference genome, correctly accessing transcripts of genes positioned in highly polymorphic genomic regions, or generating particular transcriptomes from divergent genotypes.

However, advances in sequencing technologies, also improvement of *de novo* assemblers in terms of data accuracy, and some specific issues in RNA-Seq studies, indicate that both strategies (genome-guided and *de novo*) tend to be necessary for the foreseeable future. Some works use both strategies to ensure that unique genes to a distinct cultivar (not covered in the reference genome) can also be identified and validated (e.g., Kovi, Amdahl, Alsheikh, & Rognli, 2017). Besides, the *de novo* approach is mandatory for researchers working with orphan species without a reference genome.

Recently, RNA-Seq data generated by the SGS (Illumina short reads) technique have been combined with long reads, such as those from single-molecule real time (SMRT, PacBio) sequencing, allowing higher assembly reliability, besides the identification of splice variants (e.g., Zhang et al., 2020).

10.4.4 Accessing the overall quality of the assembly

The availability of an accurate set of reference transcript sequences is critical for all downstream steps in the analysis. *De novo* assemblies are generally more susceptible to errors, but they are notably favored by paired-end reads, which improves the correct joining of exons in the same contig. Since *de novo* assembly methods are particularly more complicated to perform, they require special attention when assessing the overall assembly quality.

When good-quality reference sequences (genome/transcriptome/proteome) are available for the species studied or a closely related species, in such a case, a good practice is to compare the generated assembly with the reference sequences to evaluate the assembly's accuracy. Parameters such as the proportion of transcripts uniquely aligned to the references, average coverage per transcript, contiguity of the alignments, and the ratio of the differences observed in the alignments are good indicators of the transcriptome assembly quality.

Otherwise, when a reliable reference set is not available, the assembled transcripts' completeness can be analyzed using BUSCO (Benchmarking Universal Single-Copy Orthologs) gene set as a reference (Seppy, Manni, & Zdobnov, 2019).

10.4.5 Transcript quantification

Expression values are obtained based on the read coverage of each transcript represented in the transcriptome. When provided the reference genome, reads are mapped to the exonic regions of genes throughout the whole genome, allowing to quantify the expression of each represented gene. When a high-quality reference genome sequence is available, the genome-based approach is more appropriate for counting reads by gene rather than splice isoforms or transcripts. Without using a reference genome to estimate expression values by gene, transcript, or splice isoform, two approaches are available: alignment-based methods, when reads are aligned to the transcript sequences, and alignment-free methods, when reads are pseudo-aligned to the transcriptome to deduce the count of reads paired with each transcript. Alignment-free procedures show significant improvement in speed and memory usage compared to the alignment-based methods (Bray, Pimentel, Melsted, & Pachter, 2016).

10.4.6 Differential expression analysis

A gene expression analysis at a genome-scale, revealing genes or transcripts presenting relevant differences in gene expression profiles concerning two experimental conditions, is the first common RNA-Seq application. The approach helps to understand the molecular basis of phenotypic variations involving functional, developmental, or stress responses.

Before any comparison, it is essential to properly normalize the count data, considering factors that may impact the count values, such as transcript length, read depth per sample, and sequencing bias as GC-content (guanine-cytosine content), if necessary. The normalization of counting data aims to minimize the systematic effects of these factors. For example, assuming that longer transcripts add a larger count of reads mapped along the sequence, compared to shorter transcripts at the same level of expression, normalization by transcript length would be necessary to compare expression between different genes within the same sample, but not to compare the expression of the same gene in different samples.

Methods such as RPKM (Reads Per Kilobase of exon sequence, per Million reads mapped), FPKM (Fragments Per Kilobase of exon sequence, per Million fragments mapped), and TPM (Transcripts Per Million) consider the read length normalization and generally performed during the counting process. However, differential expression analysis is based on comparing expression values between samples. For this, it is more appropriate to use a normalization method designed to correct biases in the read depth per sample, like the TMM (Trimmed Mean of M-value), available in edgeR (Robinson, McCarthy, & Smyth, 2010), and RLE (Relative Log Estimate), available in DESeq2/DESeq (Love, Huber, & Anders, 2014).

If properly normalized at the different expression levels, the data could be centered in zero and spread evenly along the y -axis in the MA plot. This plot is a scatter plot of two experimental groups with the transcript abundance differences on the y -axis and the average of normalized expression counts on the x -axis, both on the logarithmic scale with base two. Additionally, a GC-content normalization could be done by performing EDAseq (Risso, Schwartz, Sherlock, & Dudoit, 2011).

The most straightforward approach for measuring expression changes by comparing different conditions is the fold change ratio between the average counts in each condition, usually expressed as a base two logarithm. Fold change units describe the ratio between two count values, but not exactly the difference. Methods based on statistical tests, considering the variability in expression levels from replicates, indicate whether the gene or transcript is differentially expressed or not, comparing two experimental situations.

10.4.7 Annotation and functional analysis

After the differential expression analysis, it is fundamental to explore how differentially expressed genes are related to the biologic context under study. For this purpose, it is essential to connect the observed gene expression profiles with

the currently available knowledge about the function of genes and proteins, including potential protein families and functional classification resources (InterPro, Pfam, HMMER), biochemical metabolic pathways, and gene ontology, helping to explain specific molecular biology or functional genomics aspects involved. The goal is to understand how the components (e.g., differentially expressed genes) of the molecular system are involved in the process that triggers the observed phenotype.

10.5 Functional genomics

As genomic and transcriptomic analyses increase information about genes and their regulation under experimental conditions, including abiotic stress (Wang et al., 2018; Zhang et al., 2017), also proteomics studies presented good predictive models of expression, describing, in addition to protein abundances, protein–protein interaction (PPI)/metabolite interactions, and inferences about metabolic pathways (Luan et al., 2018). However, through the NGS technology expressive amount of data has been regularly increased, and together with the growing refinements of the analysis, bioinformatics was inserted into the big data scenario. Thus the suffix -omics added to biology disciplines defined particular fields of study, and besides genomics, transcriptomics, and proteomics, also the fields metabolomics, interactomics, phenomics, among others, were derived. These multiomic approaches are complementary, and they allow the structural, functional, and dynamic characterization of organisms.

While transcriptomics focuses on the RNA abundances corresponding to genes expressed in a biological sample under particular conditions, proteomics encompasses the set of proteins and their activities. In turn, metabolomics focuses on the chemical processes and the nature of metabolites produced during metabolic processes, while interactomics covers several interactions, including PPIs and their consequences, and finally, phenomics analyzes organisms based on phenotypes. All of these multiomic approaches contribute to functional genomics (Hong, Kim, Chandran, & Jung, 2019).

With the NGS data increasing its storage in public and private databases, the development of bioinformatics resources provided organization, standardization, and curation of databases (Arora et al., 2018). In general, biological databases provide users with the search and retrieval of data, applying query models in the most diverse layers of knowledge, exploring, depending on the database, resources such as DNA/RNA annotations and analyzes, protein sequences, metabolites, molecular structures, and expression profiles (Arora et al., 2018; Marr, 2018; Rao, Das, Rao, & Srinubabu, 2008).

Regarding plant genomics resources, the website portal Phytozome, the Plant Comparative Genomics platform of the Department of Energy's Joint Genome Institute (<https://phytozome.jgi.doe.gov/pz/portal.html>, accession: December 2020), hosts (v13, with genomes released since May 2019) 224 assembled and annotated genomes covering 128 *Archaeplastida* species (235 genomes from *Viridiplantae* species). The functionally annotated genes, a powerful resource, are based on protein analysis using several resources, including KOG (euKaryotic Orthologous Groups), KEGG (Kyoto Encyclopedia of Genes and Genomes), and InterPro families (Goodstein et al., 2012).

Concerning plant metabolomics, the KNApSAcK family databases (<http://www.knapsackfamily.com/KNApSAcK/>; Afendi et al., 2012) integrate metabolite-plant species databases for multifaceted plant research. Performing a search by the organism (the scientific name), it is possible to identify/select metabolites, among 53,032 compounds and 23,911 species, according to the last update of the database (11/19/2020). Another plant metabolomics resource for gene expression association with metabolite accumulation is the platform PRIME (<http://prime.psc.riken.jp/>; Akiyama et al., 2008), the website portal from RIKEN (CSRS, Center for Sustainable Resource Science, Japan), which provides tools ranging from transcriptomics to metabolomics (Sakurai et al., 2013).

About phenomics, since functional genomics studies allow the prediction of plant genes controlling desirable agronomic traits, a practical way to identify gene functions is to analyze phenotypic differences expressed by accessions or compare them with wild-type plants. The NGS technologies enabled, for example, to characterize 91,513 mutations associated with 32,307 rice genes (Hong et al., 2019), providing information indexed in the cured databases. These mutations are powerful resources qualifying functional characterization of genes and respective genotypes/phenotypes.

The constant evolution of bioinformatics resources increases the data's confidence and facilitates functional annotation of candidate genes. In this regard, a multiomic data analysis by integrative approaches provides the proposition of new hypotheses in a biological context.

10.6 Final considerations

In the early days of transcriptomics and first-generation DNA sequencing, even a limited number of partial transcripts (e.g., ESTs) already indicated the potential of that approach. However, such studies lacked depth and involved

experimental designs that did not provide robust statistical support to assess the global gene expression. Since then, several techniques (with open or closed architecture) have been developed. Despite many possibilities, the RNA-Seq performed with the NGS technique is undoubtedly the most accurate and disseminated. Such development also benefited from the ability to generate many reads at progressively more affordable costs. However, the presented RNA-Seq overview covering cultivated plants exposed to abiotic stress pointed out that most of the identified manuscripts (covering the last 5 years) did not report any physiological data of those plants exposed to the abiotic stress. Also, some validation steps of the *in silico* differential expression detected for selected genes, usually by RT-qPCR assay, were performed with only one reference gene, sometimes not following the MIQE guidelines.

The overall RNA-Seq analysis workflow suggested a genome-guided analysis of the RNA-Seq data instead of the *de novo* assembly strategy when a high-quality reference genome of the organism is available. Probably, the number of differentially expressed genes is higher in *de novo* assemblies due to the lack of strand-specific information. Studies using (and comparing) both strategies are still scarce. *De novo* assembly strategy is also justified considering that different accessions may present exclusive gene regions worthy of validation by qPCR (in this case, essential). Besides, both methods identify many similar candidate genes, thus demonstrating the potential of the *de novo* method of capturing gene candidates, even in the absence of a reference genome. There is an increasing tendency for hybrid assembly approaches, that is, associating short cDNA sequencing (e.g., Illumina) to long (e.g., SMRT) genome sequencing, considered a promising strategy for researchers working with orphan crops without a reference genome.

Concerning the complexity and size of plant genomes, bioinformatics approaches currently possible (and recommended) raise plant transcriptomics at the level of big data, providing massive information to be validated and understood by the scientific community. It should be noted that accurate bioinformatics analysis requires databases continually curated and updated. In this context, plant databases covering genomic and transcriptomic data are well represented compared with other omics databases, such as lipidome, ionome, or phenome, currently less represented. Thus genomic/transcriptomic analyses are favored due to the higher number of cured data and databases. These databases stand out among the arsenal of bioinformatics resources, and since crop productivity depends on functional genes, those databases with appropriate approaches may assist researchers in the selection of genes associated with desirable agronomic traits. Also, integrative approaches exploring multiomic databases will facilitate machine learning in identifying gene networks for desirable agronomic traits, helping to discover new genes and related metabolic subpathways. Additionally, with the development of tools or applications, including target gene editing (CRISPR technology; Cong *et al.*, 2013), the biotechnological manipulation of genes could improve key agronomic traits.

Acknowledgments

The authors are grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 311894/2017–8), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) for financial support and fellowships.

References

- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., et al. (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science (New York, N.Y.)*, 252, 1651–1656. Available from <https://doi.org/10.1126/science.2047873>.
- Afendi, F. M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., et al. (2012). KNApSACk family databases: Integrated metabolite-plant species databases for multifaceted plant research. *Plant & Cell Physiology*, 53, e1. Available from <https://doi.org/10.1093/pcp/pcr165>.
- Akiyama, K., Chikayama, E., Yuasa, H., Shimada, Y., Tohge, T., Shinozaki, K., et al. (2008). PRIME: A web site that assembles tools for metabolomics and transcriptomics. *In Silico Biology*, 8, 339–345.
- Arora, D., Budhlakoti, N., Mishra, D. C., Chaturvedi, K. K., Kumar, S., Pradhan, A., et al. (2018). Use of Bioinformatics in crop improvement. *An International Journal of Biological Sciences*, 8, 88–93. Available from <https://doi.org/10.5958/2322-0996.2018.00001.7>.
- Bass, A. J., Robinson, D. G., & Storey, J. D. (2019). Determining sufficient sequencing depth in RNA-seq differential expression studies. *Genomics*. Available from <https://doi.org/10.1101/635623>.
- Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34, 525–527. Available from <https://doi.org/10.1038/nbt.3519>.
- Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., et al. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, 55, 611–622. Available from <https://doi.org/10.1373/clinchem.2008.112797>.

- Chaves, M. M., Flexas, J., & Pinheiro, C. (2009). Photosynthesis under drought and salt stress: Regulation mechanisms from whole plant to cell. *Annals of Botany*, *103*, 551–560. Available from <https://doi.org/10.1093/aob/mcn125>.
- Chung, P. J., Jung, H., Jeong, D.-H., Ha, S.-H., Choi, Y. Do, & Kim, J.-K. (2016). Transcriptome profiling of drought responsive noncoding RNAs and their target genes in rice. *BMC Genomics*, *17*, 563. Available from <https://doi.org/10.1186/s12864-016-2997-3>.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. Available from <https://doi.org/10.1186/s13059-016-0881-8>.
- Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, *339*, 819–823. Available from <https://doi.org/10.1126/science.1231143>.
- Cui, G., Chai, H., Yin, H., Yang, M., Hu, G., Guo, M., et al. (2019). Full-length transcriptome sequencing reveals the low-temperature-tolerance mechanism of *Medicago falcata* roots. *BMC Plant Biology*, *19*, 575. Available from <https://doi.org/10.1186/s12870-019-2192-1>.
- Dang, Z., Zheng, L., Wang, J., Gao, Z., Wu, S., Qi, Z., et al. (2013). Transcriptomic profiling of the salt-stress response in the wild rethorhalophyte *Reaumuria trigyna*. *BMC Genomics*, *14*, 29. Available from <https://doi.org/10.1186/1471-2164-14-29>.
- Danilevskaya, O. N., Yu, G., Meng, X., Xu, J., Stephenson, E., Estrada, S., et al. (2019). Developmental and transcriptional responses of maize to drought stress under field conditions. *Plant Direct*, *3*, e00129. Available from <https://doi.org/10.1002/pld3.129>.
- de Lima Cabral, G. A., Binneck, E., de Souza, M. C. P., da Silva, M. D., Costa Ferreira Neto, J. R., Pompelli, M. F., et al. (2020). First expressed TFome of physic nut (*Jatropha curcas* L.) after salt stimulus. *Plant Molecular Biology Reporter*, *38*, 189–208. Available from <https://doi.org/10.1007/s11105-019-01187-w>.
- Di Bella, S., La Ferlita, A., Carapezza, G., Alaimo, S., Isacchi, A., Ferro, A., et al. (2019). A benchmarking of pipelines for detecting ncRNAs from RNA-seq data. *Briefings in Bioinformatics*. Available from <https://doi.org/10.1093/bib/bbz110>.
- Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., et al. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature*, *260*, 500–507. Available from <https://doi.org/10.1038/260500a0>.
- Flicek, P., & Birney, E. (2009). Sense from sequence reads: Methods for alignment and assembly. *Nature Methods*, *6*, S6–S12. Available from <https://doi.org/10.1038/nmeth.1376>.
- Geniza, M., & Jaiswal, P. (2017). Tools for building de novo transcriptome assembly. *Current Plant Biology*, *11–12*, 41–45. Available from <https://doi.org/10.1016/j.cpb.2017.12.004>.
- Glass, D. J. (2014). *Cold Spring Harbor Experimental design for biologists* (Second edition). New York: Cold Spring Harbor Laboratory Press.
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*, D1178–D1186. Available from <https://doi.org/10.1093/nar/gkr944>.
- Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., et al. (1965). Structure of a ribonucleic acid. *Science (80-)*, *147*, 1462 LP–1461465. Available from <https://doi.org/10.1126/science.147.3664.1462>.
- Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-seq assemblers. *GigaScience*, *8*. Available from <https://doi.org/10.1093/gigascience/giz039>, giz039.
- Hong, W.-J., Kim, Y.-J., Chandran, A. K. N., & Jung, K.-H. (2019). Infrastructures of systems biology that facilitate functional genomics study in rice. *Rice (N Y)*, *12*, 15. Available from <https://doi.org/10.1186/s12284-019-0276-z>.
- Huang, B.-L., Li, X., Liu, P., Ma, L., Wu, W., Zhang, X., et al. (2019). Transcriptomic analysis of *Eruca vesicaria* subs. *sativa* lines with contrasting tolerance to polyethylene glycol-simulated drought stress. *BMC Plant Biology*, *19*, 419. Available from <https://doi.org/10.1186/s12870-019-1997-2>.
- Hübner, S., Korol, A. B., & Schmid, K. J. (2015). RNA-seq analysis identifies genes associated with differential reproductive success under drought-stress in accessions of wild barley *Hordeum spontaneum*. *BMC Plant Biology*, *15*, 134. Available from <https://doi.org/10.1186/s12870-015-0528-z>.
- Kido, E. A., Ferreira Neto, J. R., Kido, S. A. B., Pandolfi, V., & Benko-Iseppon, A. M. (2013). DeepSuperSAGE in a friendly bioinformatic approach: Identifying molecular targets responding to abiotic stress in plants. In R. K. Gaur, & P. Sharma (Eds.), *Molecular approaches in plant abiotic stress* (2013, pp. 108–126). CRC Press.
- Kido, É. A., Ferreira-Neto, J. R. C., Pandolfi, V., de Melo Souza, A. C., & Benko-Iseppon, A. M. (2016). *Drought stress tolerance in plants: Insights from transcriptomic studies*. *Drought Stress Toler. Plants* (Vol 2, pp. 153–185). Springer.
- Kovi, M. R., Amdahl, H., Alsheikh, M., & Rognli, O. A. (2017). De novo and reference transcriptome assembly of transcripts expressed during flowering provide insight into seed setting in tetraploid red clover. *Scientific Reports*, *7*, 44383. Available from <https://doi.org/10.1038/srep44383>.
- Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Saïnderichin, E., Hu, G., et al. (2018). Optimization of an RNA-seq differential gene expression analysis depending on biological replicate number and library size. *Frontiers of Plant Science*, *9*, 108. Available from <https://doi.org/10.3389/fpls.2018.00108>.
- Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., et al. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, *7*, 709–715. Available from <https://doi.org/10.1038/nmeth.1491>.
- Li, F.-D., Tong, W., Xia, E.-H., & Wei, C.-L. (2019). Optimized sequencing depth and de novo assembler for deeply reconstructing the transcriptome of the tea plant, an economically important plant species. *BMC Bioinformatics*, *20*, 553. Available from <https://doi.org/10.1186/s12859-019-3166-x>.
- Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics (Oxford, England)*, *30*, 301–304. Available from <https://doi.org/10.1093/bioinformatics/btt688>.
- Lorkowski, S., & Cullen, P. M. (2006). *Analysing Gene Expression: A Handbook of Methods, Possibilities, and Pitfalls*. John Wiley & Sons.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550. Available from <https://doi.org/10.1186/s13059-014-0550-8>.
- Luan, H., Shen, H., Pan, Y., Guo, B., Lv, C., & Xu, R. (2018). Elucidating the hypoxic stress response in barley (*Hordeum vulgare* L.) during water-logging: A proteomics approach. *Scientific Reports*, *8*, 9655. Available from <https://doi.org/10.1038/s41598-018-27726-1>.

- Ma, N., Hu, C., Wan, L., Hu, Q., Xiong, J., & Zhang, C. (2017). Strigolactones improve plant growth, photosynthesis, and alleviate oxidative stress under salinity in rapeseed (*Brassica napus* L.) by regulating gene expression. *Frontiers of Plant Science*, 8, 1671. Available from <https://doi.org/10.3389/fpls.2017.01671>.
- Ma, S., Lv, L., Meng, C., Zhou, C., Fu, J., Shen, X., et al. (2019). Genome-wide analysis of abscisic acid biosynthesis, catabolism, and signaling in sorghum bicolor under saline-alkali stress. *Biomolecules*, 9. Available from <https://doi.org/10.3390/biom9120823>.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376–380. Available from <https://doi.org/10.1038/nature03959>.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18, 1509–1517. Available from <https://doi.org/10.1101/gr.079558.108>.
- Marr, T. (2018). *Computers and DNA*. Routledge. Available from <https://doi.org/10.4324/9780429501463>.
- Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., et al. (2005). SuperSAGE. *Cellular Microbiology*, 7, 11–18. Available from <https://doi.org/10.1111/j.1462-5822.2004.00478.x>.
- Matsumura, H., Yoshida, K., Luo, S., Kimura, E., Fujibe, T., Albertyn, Z., et al. (2010). High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS One*, 5, e12010. Available from <https://doi.org/10.1371/journal.pone.0012010>.
- Molina, C., Zaman-Allah, M., Khan, F., Fatnassi, N., Horres, R., Rotter, B., et al. (2011). The salt-responsive transcriptome of chickpea roots and nodules via deepSuperSAGE. *BMC Plant Biology*, 11, 31. Available from <https://doi.org/10.1186/1471-2229-11-31>.
- Moon, K.-B., Ahn, D.-J., Park, J.-S., Jung, W. Y., Cho, H. S., Kim, H.-R., et al. (2018). Transcriptome profiling and characterization of drought-tolerant potato plant (*Solanum tuberosum* L.). *Molecules and Cells*, 41, 979–992. Available from <https://doi.org/10.14348/molcells.2018.0312>.
- Morgil, H., Tardu, M., Cevahir, G., & Kavakli, İ. H. (2019). Comparative RNA-seq analysis of the drought-sensitive lentil (*Lens culinaris*) root and leaf under short- and long-term water deficits. *Functional & Integrative Genomics*, 19, 715–727. Available from <https://doi.org/10.1007/s10142-019-00675-2>.
- Mousavi, S., Alisoltani, A., Shiran, B., Fallahi, H., Ebrahime, E., Imani, A., et al. (2014). De novo transcriptome assembly and comparative analysis of differentially expressed genes in *Prunus dulcis* Mill. in response to freezing stress. *PLoS One*, 9, e104541. Available from <https://doi.org/10.1371/journal.pone.0104541>.
- Murphy, D. (2002). Gene expression studies using microarrays: Principles, problems, and prospects. *Advances in Physiology Education*, 26, 256–270. Available from <https://doi.org/10.1152/advan.00043.2002>.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)* (80-), 320, 1344–1349. Available from <https://doi.org/10.1126/science.1158441>.
- Peng, Z., He, S., Gong, W., Sun, J., Pan, Z., Xu, F., et al. (2014). Comprehensive analysis of differentially expressed genes and transcriptional regulation induced by salt stress in two contrasting cotton genotypes. *BMC Genomics*, 15, 760. Available from <https://doi.org/10.1186/1471-2164-15-760>.
- Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge, UK ; New York: Cambridge University Press.
- Rao, V. S., Das, S. K., Rao, V. J., & Srinubabu, G. (2008). Recent developments in life sciences research: Role of bioinformatics. *African Journal of Biotechnology*, 7. Available from <https://doi.org/10.5897/AJB07.899>.
- Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinformatics*, 12, 480. Available from <https://doi.org/10.1186/1471-2105-12-480>.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26, 139–140. Available from <https://doi.org/10.1093/bioinformatics/btp616>.
- Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., et al. (2002). Using the transcriptome to annotate the genome. *Nature Biotechnology*, 20, 508–512. Available from <https://doi.org/10.1038/nbt0502-508>.
- Sakurai, T., Yamada, Y., Sawada, Y., Matsuda, F., Akiyama, K., Shinozaki, K., et al. (2013). PRIME update: Innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant & Cell Physiology*, 54, e5. Available from <https://doi.org/10.1093/pcp/pcs184>.
- Sánchez-Romera, B., & Aroca, R. (2020). *Plant roots—the hidden half for investigating salt and drought stress responses and tolerance. Salt and drought stress tolerance in plants* (pp. 137–175). Springer. Available from https://doi.org/10.1007/978-3-030-40277-8_6.
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270, 467–470. Available from <https://doi.org/10.1126/science.270.5235.467>.
- Seppy, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. New York In M. Kollmar (Ed.), *Gene Prediction* (vol. 1962, pp. 227–245). New York, NY: Springer. Available from https://doi.org/10.1007/978-1-4939-9173-0_14.
- Shahid, S. A., Zaman, M., & Heng, L. (2018). *Soil salinity: Historical perspectives and a world overview of the problem. Guideline for salinity assessment, mitigation and adaptation using nuclear and related techniques* (pp. 43–53). Springer.
- Shiferaw, B., Prasanna, B. M., Hellin, J., & Bänziger, M. (2011). Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security*, 3, 307.
- Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews. Genetics*, 20, 631–656. Available from <https://doi.org/10.1038/s41576-019-0150-2>.
- Swann, A. L. S. (2018). Plants and drought in a changing climate. *Current Climate Change Reports*, 4, 192–201. Available from <https://doi.org/10.1007/s40641-018-0097-y>.
- Tang, Y., Wang, J., Bao, X., Liang, M., Lou, H., Zhao, J., et al. (2019). Genome-wide identification and expression profile of HD-ZIP genes in physic nut and functional analysis of the JcHDZ16 gene in transgenic rice. *BMC Plant Biology*, 19, 298. Available from <https://doi.org/10.1186/s12870-019-1920-x>.

- Temin, H. M., & Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of *Rous sarcoma virus*. *Nature*, 226, 1211–1213. Available from <https://doi.org/10.1038/2261211a0>.
- Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science (New York, N.Y.)*, 270, 484–487. Available from <https://doi.org/10.1126/science.270.5235.484>.
- Wan, X. L., Zhou, Q., Wang, Y. Y., Wang, W. E., Bao, M. Z., & Zhang, J. W. (2015). Identification of heat-responsive genes in carnation (*Dianthus caryophyllus* L.) by RNA-seq. *Frontiers of Plant Science*, 6, 519. Available from <https://doi.org/10.3389/fpls.2015.00519>.
- Wang, L., Li, D., Zhang, Y., Gao, Y., Yu, J., Wei, X., et al. (2016). Tolerant and susceptible sesame genotypes reveal waterlogging stress response patterns. *PLoS One*, 11, e0149912. Available from <https://doi.org/10.1371/journal.pone.0149912>.
- Wang, P., Su, L., Gao, H., Jiang, X., Wu, X., Li, Y., et al. (2018). Genome-wide characterization of bHLH genes in grape and analysis of their potential relevance to abiotic stress tolerance and secondary metabolite biosynthesis. *Frontiers of Plant Science*, 9, 64. Available from <https://doi.org/10.3389/fpls.2018.00064>.
- Wang, R., Mei, Y., Xu, L., Zhu, X., Wang, Y., Guo, J., et al. (2018). Genome-wide characterization of differentially expressed genes provides insights into regulatory network of heat stress response in radish (*Raphanus sativus* L.). *Functional & Integrative Genomics*, 18, 225–239. Available from <https://doi.org/10.1007/s10142-017-0587-3>.
- Wang, W., Xin, H., Wang, M., Ma, Q., Wang, L., Kaleri, N. A., et al. (2016). Transcriptomic analysis reveals the molecular mechanisms of drought-stress-induced decreases in *Camellia sinensis* leaf quality. *Frontiers of Plant Science*, 7, 385. Available from <https://doi.org/10.3389/fpls.2016.00385>.
- Wu, H., Wang, C., & Wu, Z. (2015). PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics (Oxford, England)*, 31, 233–241. Available from <https://doi.org/10.1093/bioinformatics/btu640>.
- Wu, J., Zhao, Q., Wu, G., Yuan, H., Ma, Y., Lin, H., et al. (2019). Comprehensive analysis of differentially expressed Unigenes under NaCl stress in flax (*Linum usitatissimum* L.) using RNA-seq. *International Journal of Molecular Sciences*, 20. Available from <https://doi.org/10.3390/ijms20020369>.
- Wu, L., Taohua, Z., Gui, W., Xu, L., Li, J., & Ding, Y. (2015). Five pectinase gene expressions highly responding to heat stress in rice floral organs revealed by RNA-seq analysis. *Biochemical and Biophysical Research Communications*, 463, 407–413. Available from <https://doi.org/10.1016/j.bbrc.2015.05.085>.
- Wu, P., Cogill, S., Qiu, Y., Li, Z., Zhou, M., Hu, Q., et al. (2020). Comparative transcriptome profiling provides insights into plant salt tolerance in seashore paspalum (*Paspalum vaginatum*). *BMC Genomics*, 21, 131. Available from <https://doi.org/10.1186/s12864-020-6508-1>.
- Wu, Z., & Wu, H. (2016). Experimental design and power calculation for RNA-seq experiments, New York In E. Mathé, & S. Davis (Eds.), *Statistical Genomics* (vol. 1418, pp. 379–390). New York, NY: Springer. Available from https://doi.org/10.1007/978-1-4939-3578-9_18.
- Xu, J., Zhang, M., Liu, G., Yang, X., & Hou, X. (2016). Comparative transcriptome profiling of chilling stress responsiveness in grafted watermelon seedlings. *Plant Physiology and Biochemistry*, 109, 561–570. Available from <https://doi.org/10.1016/j.plaphy.2016.11.002>.
- Yang, T., & Huang, X.-S. (2018). Deep sequencing-based characterization of transcriptome of *Pyrus ussuriensis* in response to cold stress. *Gene*, 661, 109–118. Available from <https://doi.org/10.1016/j.gene.2018.03.067>.
- Zawada, A. M., Rogacev, K. S., Müller, S., Rotter, B., Winter, P., Fliser, D., et al. (2014). Massive analysis of cDNA Ends (MACE) and miRNA expression profiling identifies proatherogenic pathways in chronic kidney disease. *Epigenetics: Official Journal of the DNA Methylation Society*, 9, 161–172. Available from <https://doi.org/10.4161/epi.26931>.
- Zenda, T., Liu, S., Wang, X., Liu, G., Jin, H., Dong, A., et al. (2019). Key maize drought-responsive genes and pathways revealed by comparative transcriptome and physiological analyses of contrasting inbred lines. *International Journal of Molecular Sciences*, 20. Available from <https://doi.org/10.3390/ijms20061268>.
- Zhang, A., Han, D., Wang, Y., Mu, H., Zhang, T., Yan, X., et al. (2018). Transcriptomic and proteomic feature of salt stress-regulated network in Jerusalem artichoke (*Helianthus tuberosus* L.) root based on de novo assembly sequencing analysis. *Planta*, 247, 715–732. Available from <https://doi.org/10.1007/s00425-017-2818-1>.
- Zhang, D., Li, W., Chen, Z.-J., Wei, F.-G., Liu, Y.-L., & Gao, L.-Z. (2020). SMRT- and Illumina-based RNA-seq analyses unveil the ginsenoside biosynthesis and transcriptomic complexity in *Panax notoginseng*. *Scientific Reports*, 10, 15310. Available from <https://doi.org/10.1038/s41598-020-72291-1>.
- Zhang, L., Li, X., Ma, B., Gao, Q., Du, H., Han, Y., et al. (2017). The Tartary buckwheat genome provides insights into rutin biosynthesis and abiotic stress tolerance. *Molecular Plant Pathology*, 10, 1224–1237. Available from <https://doi.org/10.1016/j.molp.2017.08.013>.
- Zhang, N., Liu, B., Ma, C., Zhang, G., Chang, J., Si, H., et al. (2014). Transcriptome characterization and sequencing-based identification of drought-responsive genes in potato. *Molecular Biology Reports*, 41, 505–517. Available from <https://doi.org/10.1007/s11033-013-2886-7>.
- Zhao, S., Li, C.-I., Guo, Y., Sheng, Q., & Shyr, Y. (2018). RnaSeqSampleSize: Real data based sample size estimation for RNA sequencing. *BMC Bioinformatics*, 19, 191. Available from <https://doi.org/10.1186/s12859-018-2191-5>.
- Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., et al. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16, 675. Available from <https://doi.org/10.1186/s12864-015-1876-7>.
- Zhao, X., Li, C., Wan, S., Zhang, T., Yan, C., & Shan, S. (2018). Transcriptomic analysis and discovery of genes in the response of *Arachis hypogaea* to drought stress. *Molecular Biology Reports*, 45, 119–131. Available from <https://doi.org/10.1007/s11033-018-4145-4>.
- Zhao, Y., Wei, X., Ji, X., & Ma, W. (2019). Endogenous NO-mediated transcripts involved in photosynthesis and carbohydrate metabolism in alfalfa (*Medicago sativa* L.) seedlings under drought stress. *Plant Physiology and Biochemistry*, 141, 456–465. Available from <https://doi.org/10.1016/j.plaphy.2019.06.023>.

This page intentionally left blank

Identification of novel RNAs in plants with the help of next-generation sequencing technologies

Aditya Narayan¹ and Shailesh Kumar²

¹University of Virginia, Charlottesville, VA, United States, ²Bioinformatics Laboratory, National Institute of Plant Genome Research (NIPGR), Delhi, New Delhi, India

11.1 Introduction

RNA has largely been regarded simply as the intermediate between DNA and protein products. However, the last several years have led to the identification of a veritable cornucopia of novel roles including, but not limited to, genome stability changes, expression modification, and physiological changes. Collectively, these functions speak to a deep and not fully discovered biological significance in plants (Shin & Shin, 2016). Such scientific breakthroughs have been facilitated by the integration of novel technologies such as high-throughput RNA-seq methodologies into standard research practice. In effect, this speaks to the need to annotate, characterize, and understand RNA types found within the plant transcriptome (Morgado & Johannes, 2019). Accordingly, this chapter will serve as a primer for the application of novel databases, tools, software packages, data science techniques, and the like to the plant genome.

11.1.1 Noncoding RNA classes in plants

There are a range of essential noncoding RNAs (ncRNA), including transfer RNA (tRNA), small-interfering RNA (siRNA), ribosomal RNA (rRNA), small nucleolar RNA, tRNA-derived small RNA (tsRNA), microRNA (miRNA), heterochromatic siRNA (hc-siRNA), natural antisense siRNA (NAT-siRNA), phased siRNA (phasiRNA), trans-acting siRNA (tasiRNA), and several others, in which each possesses key regulatory functions (Bailey-Serres, Zhai, & Seki, 2020; Borges & Martienssen, 2015; Morgado & Johannes, 2019; Szakonyi, Confraria, Valerio, Duque, & Staiger, 2019). These molecules, which can be categorized as short, long, linear, or circular ncRNA, vary enormously with respect to function, structure, and distribution (Cao, Wahlestedt, & Kapranov, 2018; Liao, Li, Cui, & Zheng, 2018; Signal, Gloss, & Dinger, 2016).

The abovementioned ncRNA classes may be broadly categorized as small RNA (sRNA), are created through the action of Dicer-like (DCL) proteins, and are most frequently found in transcriptomes with length ranging from 20 to 24 nt. These small RNAs are formed from two groups of precursor molecules: those with hairpin structures (hpRNA), and double-stranded RNA (Axtell, 2013). The former encompasses miRNA and non-miRNA while the latter describes highly abundant siRNA and the associated forms of siRNA (Borges & Martienssen, 2015).

11.2 Small RNA

11.2.1 MicroRNA

miRNAs are a class of noncoding small RNA transcribed from adjacent miRNA genes as a polycistron or monocistron by the action of RNA polymerase II. After splicing to remove introns, polyadenylation, and addition of the 5' cap, the final product is a 20- to 24-nt molecule which folds into hairpin-like structures. At this stage the molecule is classified

as a primary miRNA. They subsequently undergo nuclear cleavage by DCL1 into precursor miRNA (pre-miRNA) and are cleaved again in the cytoplasm to produce a mature miRNA duplex. The duplex unwinds and the mature miRNA may then take on a functional role by assembling with Argonaute proteins to create the RNA-induced silencing complex, or RISC, which guides cleavage or repression of target mRNA by pairing with complementary targets (Jones-Rhoades, Bartel, & Bartel, 2006; Lee, Jeon, Lee, Kim, & Kim, 2002; Wang, Mei, & Ren, 2019). The miRNAs are also capable of regulating protein-coding mRNA or other ncRNAs through repression due to cleavage or inhibition of translation (Axtell, 2013).

It is also briefly worth discussing an alternative form of miRNA: the isomiRNA or isomiR. The identification of such variants has resulted from advances in next-generation and deep sequencing which may be applied for the identification of substitutions and edits at the termini. Variations may be classified as trimming variants, in which the cleavage site of the DCLs is changed, nucleotide addition variants, and nucleotide substitution variants. This class may be identified through software algorithms trained to identify mapping mismatches, modifications such as adenylation, and by analysis of origin (template or nontemplate). Analogous to miRNA, the isomiRNA function as regulatory molecules (Neilsen, Goodall, & Bracken, 2012; Sablok, Srivastva, Suprasanna, Baev, & Ralph, 2015).

To identify miRNA, it is necessary to screen the genome for the appropriate loci. These consist of hairpin loop sequences. However, issues may arise in this process as miRNA loci often have high false-positive rates of identification. The associated algorithm also mandates a great deal of prior information to more accurately identify these loci and broadly applies RNA sequencing data in tandem with genomic sequencing approaches (Bortolomeazzi, Gaffo, & Bortoluzzi, 2017). RNA-seq or sRNA seq kits may be used to collect the necessary transcriptomic data and it is subsequently cleaned to discard unrelated RNA class information. This dataset may then be aligned to a reference genome, though if this is not possible to a dearth of extant genomes there exist alternative software tools that may be applied to the task. Bowtie or BWA accepts mRNA data as well as Expressed Sequence Tag (ESTs) (short DNA sequences generated from cDNA clones which may be used to identify coding regions of DNA), Genome Survey Sequences (analogous to ESTs but are genomic in origin), or the genomes of closely related species. To effectively validate the identification of hairpin precursor sequences, it is necessary to evaluate it based on known pathway characteristics, including secondary structure stability, loop structure, and identification of clustered sequences illustrating sites leading to the creation of miRNA–miRNA* duplexes. Mature miRNA may be detected in precursors based on relative abundance (Ma, Tang, Qin, & Meng, 2015). It is also possible to detect clusters of reads, or block groups, which exist on the same strand and have similar start/stop positions, corresponding to miRNA and miRNA*. These clusters are profiled based on similarity, read length, and location. However, this read profiling is limited for the purpose of detecting novel miRNA. Finally, it is possible to make predictions by assembling sRNA into contigs which are then filtered and used to create candidate miRNA duplexes. These candidates are examined for features, including length and paired bases relative to unpaired bases. This approach is particularly useful for those species without a reference genome as it applies machine learning algorithms using model species to create predictions in nonmodel species (Berezikov, Cuppen, & Plasterk, 2006).

Ultimately, a number of tools have become available for miRNA loci annotations based on size, abundance of miRNA and miRNA*, and the secondary structures of hairpin precursors. Each approach offers variable efficacy for different settings and thus may generate conflicting results (Gomes et al., 2013; Mendes, Freitas, & Sagot, 2009). False-positive predictions are of particular concern, and several studies have been published describing optimal criteria for annotation. Several tools seek to strictly adhere to these guidelines such as ShortStack and miR-PREFer (Axtell, 2013; Lei & Sun, 2014). Issues also arise with the fact that sRNA-seq is generally biased toward certain miRNA. It is possible to differentiate between miRNA and other sRNAs through RNA-dependent RNA polymerase (RDR) mutants as this enzyme group is not involved in miRNA synthesis (Kozomara & Griffiths-Jones, 2014).

In addition to annotation tools, recent increases in miRNA data have necessarily led to the need to establish miRNA databases for plants and other organisms. One such example is miRBase, a database that uses standard names for miRNA families, functions using guidelines to prevent incorrect annotations, and allows for searches to be performed on genomic loci, pre-miRNA, and mature miRNA (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006; Kozomara, Birgaoanu, & Griffiths-Jones, 2019). While an incredible resource, it is necessary to acknowledge that there are miRNAs present in the database which have minimal evidence supporting their existence, fail to meet standard annotation criteria, or are degraded fragments. This may largely be attributed to the presence of siRNA which interferes with efforts to sequence miRNA in a sample given their relatively similar size. Such concerns have led the research community to adopting the usage of pooled sRNA-seq libraries. These account for genomic loci which create miRNA at lower levels by applying stringent parameters. Examples of such tools include sRNAanno and PmiREN (Chen et al., 2019; Guo et al., 2019). Further, for plant-specific studies, sRNA-seq data are available in databases such as the Sequence Read Archive, Cereal small RNA database, Arabidopsis Small RNA Database, and the like. These

repositories are specific to plants or particular plant species (Feng et al., 2020; Johnson, Bowman, Adai, Vance, & Sundaresan, 2007; Leinonen, Sugawara, & Shumway, 2011). A collection of some of the most commonly cited tools and databases employed for plant miRNA annotation is found in Table 11.1.

11.2.2 Small-interfering RNA

Long dsRNAs form through hybridization of sense and antisense transcripts, hybridization of RNA with complementary sequences, RDRs, or folding of inverted repeat sequences. Precursors of siRNA are transcribed by Pol II and are formed from transposable elements, noncoding loci, and protein-coding genes. These protein-coding genes are subject to RDRs

TABLE 11.1 Most commonly cited tools and databases employed for plant miRNA annotation.

Annotation tools				
Tool	Language	Platform	Description	URL
MiReNA (2010) v2.0	C, Bash, Python, Perl	Linux	A genome-wide search algorithm that looks for miRNA sequences by exploring a multidimensional space defined by only five parameters. It validates pre-miRNAs with high sensitivity and specificity, detects new miRNAs by homology or deep-sequencing data.	http://www.lcqb.upmc.fr/mirena/index.html
PIPmiR (2011)	Java	Any platform	PIPmiR is an algorithm to identify novel plant miRNA genes using a combination of deep sequencing data and genomic characteristics. This algorithm can be used as a full pipeline, a classifier, or a precursor sequence predictor in plants.	https://bioconda.github.io/recipes/pipmir/README.html
mirDeepFinder (2012)	Perl	Linux	This software package developed to identify and functionally characterize plant miRNAs and their sequence targets from sRNA datasets obtained from deep sequencing. Functions include preprocessing of raw data, identifying miRNAs, classifying novel miRNAs, miRNA expression profile production, predicting miRNA targets, and gene pathway/network analysis.	NA
CAP-miRSeq (2014)	Perl, Python, R, Bash	Linux	This tool is an analysis pipeline for miRNA which integrates read preprocessing, alignment, mature/precursor/novel miRNA qualification, and variant detection in miRNA coding region.	http://bioinformaticstools.mayo.edu/research/cap-mirseq/
mirTools 2.0 (2013)	Perl	Linux	This offers a range of functions including, but not limited to detection and profiling of various ncRNA, identification of miRNA-targeted genes, annotating by function, and taxonomic profiling.	http://www.wzgenomics.cn/Mr2_dev/
isomiRex (2013)	Perl	Linux	isomiRex is an open-access web platform to identify isomiRs and to offer graphical visualization of the differentially expressed miRNAs.	http://bioinfo1.uniplovdiv.bg/isomiRex/
Databases				
Database	Description			URL
miRbase (2018) v22.1	This database consists of published miRNA sequences and annotation, including hairpin precursors, mature miRNA sequence, and location information across >80 plant species. It also provides microRNA functional information, related literature and provides nomenclature rules.			http://www.mirbase.org/
CSRDB (2007)	CSRDB contains maize and rice sRNA sequences developed through high-throughput pyrosequencing.			http://sundarlab.ucdavis.edu/smrnas/
mirEX (2012) v2.0	Holds details on 461 miRNAs from 3 plants (<i>Arabidopsis thaliana</i> , <i>Hordeum vulgare</i> , <i>Pellia endiviifolia</i>) and provides a platform for comparative analysis of pri-miRNA across plant species.			http://www.combio.pl/mirex2

and processed by DCL2 and DCL4 to create siRNA. Such siRNAs are involved in gene silencing and can also induce methylation of target sequences (Kamthan, Chaudhuri, Kamthan, & Datta, 2015; Wu et al., 2010). A significant natural application of siRNA is in viral defense pathways. Viral-derived siRNAs are capable of defending host cells against plant viruses and may also be used to identify known and novel plant viruses (Vivek, Zahra, & Kumar, 2019). Types of siRNA include, but are not limited to, hc-siRNA, tasiRNA, phasiRNA, NAT-siRNA, and tsRNA (Guo, Liu, Smith, Liang, & Wang, 2016).

11.2.3 Heterochromatic small-interfering RNA

Hc-siRNAs are derived from repetitive transposon-associated, intergenic regions. Hc-siRNA may function in RNA-directed DNA methylation by targeting ncRNA transcripts associated with chromatin. They are roughly 24 nt in length and originate from a dsRNA precursor of 30- to 50-nt length. The 24-nt length must be noted, given that it allows for researchers to distinguish between these and other sRNAs which are frequently 21–22 nt in length (Axtell, 2013; Huang, Wang, Hu, Hamby, & Jin, 2019; Zheng, Wang, Wu, Ding, & Fei, 2015). hc-siRNA is produced in cells from a pathway requiring RNA polymerase IV, RDR2, DCL3, and AGO4 (Deng, Muhammad, Cao, & Wu, 2018). Despite knowledge of these components, however, hc-siRNA biogenesis is poorly understood and as a result, annotation of loci producing this class of siRNA is sparse. Thus there are few no tools available for data scientists to detect, predict, or explore hc-siRNA. The database, sRNAanno, does include information on hc-siRNA in plant species based on 23–24 nt abundance (Chen et al., 2019).

11.2.4 Phased small-interfering RNA and trans-acting small-interfering RNA

PhasiRNAs are produced in a “phased” pathway in which primary transcripts are cleaved into 21-, 22-, or 24-nt length starting at a specified terminus. This initial site is determined by miRNA or siRNA cleavage from the PHAS loci which is processed to dsRNA. Phasing refers to the successive processing of dsRNA in sequence. The first cleavage site indicates the start of the first phasiRNA. After processing the phasiRNA takes on a functional role through association with AGO4 to degrade mRNA or silence transcripts. phasiRNA may also be created from long inverted repeats. tasiRNAs are a subpopulation of phasiRNA which may be produced from a tasiRNA gene which produces a noncoding transcript that induces mRNA cleavage in trans (Fei, Xia, & Meyers, 2013; Komiya, 2017; Wu, Chen, & Tian, 2017).

These siRNA classes are identified by the aforementioned phasing pattern and are broadly categorized as secondary siRNA. They may be detected by observing fixed-length in-phase and out-of-phase read accumulation around a cleavage start position. They are then assessed by ranking the potential phased arrangements. tasiRNA loci may be distinguished by observing mapping locations of sRNA reads to phasi or tasiRNA loci relative to start positions of cleavage and the target transcript (Chen, Li, & Wu, 2007; Guo, Qu, & Jin, 2015; Kakrana et al., 2017; Zheng, Wang, & Sunkar, 2014).

With respect to available tools and databases, the only active database with siRNA annotations for these classes is the tasiRNAdb (Zhang, Li, Zhu, Zhang, & Fang, 2013). Issues arise frequently in annotating these classes due to false positives from highly expressed 24-nt sRNAs passing as phasiRNA. Additional algorithms are likely necessary to correct this issue as well as to explore the potential tissue-specific expression of phasiRNA (Carbonell, 2019).

11.2.5 Natural antisense-small-interfering RNA

NAT-siRNAs are produced from dsRNA precursors but differ in that the dsRNAs of other siRNA types rely on RDR. NAT-siRNA may be broken into “cis” or “trans” groups. cis-NAT-siRNA is formed from sRNA transcribed from opposite strands of the same genomic loci which anneal to create the dsRNA. dsRNA precursors that form from nonoverlapping genes and thus have mismatches in the annealed regions are classified as trans-NAT-siRNA. Nat-siRNAs have been reported in plants, though the cis form has been identified in greater abundance (Zhang, Xia, & Lii, 2012).

NAT-siRNA has been shown to perform transcriptional interference at the RNA level as well as epigenetic modifications. However, there has been little research conducted on this particular class. As with other sRNAs, the lack of references makes classification, identification, or prediction difficult and speaks to the need for further work in this space. Current pipelines include NATpipe and NATpare (Thody, Folkes, & Moulton, 2020; Yu, Meng, Zuo, Xue, & Wang, 2016).

11.2.6 Transfer RNA–derived small RNA

tRNA fragments with functional roles represent a novel class of ncRNA which exist in plants and differ in marked ways from traditional tRNA. This particular class of RNA is produced through endonucleases acting on pre- and mature tRNA and is believed to possess functions analogous to other interfering RNA. These endonucleases include DCLs, though recent studies have validated a DCL-independent pathway as well. Broadly, the biogenesis of this class of RNA is poorly understood. tsRNAs are roughly 15–42nt in length (Li, Xu, & Sheng, 2018; Martinez, Choudury, & Slotkin, 2017).

tsRNAs are a diverse group of RNA and are classified as either tRNA halves (31–40 nt) or tRNA-derived fragments based on their size and loci in the tRNA itself (Zhu, Ge, Li, Shen, & Guo, 2019). Cleavage at the anticodon loop of a mature tRNA molecule creates tRNA halves when under stress. These, in turn, are further divided into two groups based on whether they maintain the 5' or 3' group in their structure. tRNA fragments are defined by mapping to both pre- and/or mature tRNA transcripts and may be further grouped into trF-5 (derived from 5' end of tRNA and are 14–30 nt), trF-3 (derive from 3' end of the tRNA and are 18–22 nt), and trF-1 (derived from the 3' end of pre-tRNA by RNase Z). Further diversity stems from the existence of stress-induced tRNA fragments, iso-acceptors and iso-decoders, and modifications to tRNA. Ultimately, further research is needed to properly classify this and name tsRNA classes (Alves et al., 2017; Loss-Morais, Waterhouse, & Margis, 2013; Xie et al., 2020).

Using sRNA-seq data mapped to annotated tRNA genes, it becomes possible to identify tsRNA without any knowledge of the sequence of interest. However, several issues arise in the mapping process. In particular, mapping sRNA reads to tRNA genomic loci (the tRNA space) is difficult for a number of reasons, including incomplete tRNA annotation, the fact that sRNA reads may miss tRNA produced by splicing intron regions in pre-tRNA, that the canonical CCA sequence on the 3' end of tsRNA may appear in other sequences, the overlap in size of miRNA and tsRNA, the presence of 5' phosphates and 3' hydroxyls in both tRFs and miRNA, and significant sequence modifications (Kumar, Kuscu, & Dutta, 2016).

There are a limited number of tools and databases available for studying tsRNAs. This is in part due to the fact that tRNAs are incredibly abundant and thus degraded tRNA is difficult to distinguish from functional tsRNA. Several annotation and predictive tools exist, though few are tailored for plant tsRNAs. tRex and PtRFdb are databases relevant to plant RNAs specifically (Kumar, Mudunuri, Anaya, Dutta; Thompson et al., 2018). A collection of phasiRNA, tasiRNA, and NAT-siRNA predictive tools and relevant databases are presented in Table 11.2.

11.3 Long noncoding RNA

Long ncRNAs (lncRNAs) are RNA without coding potential which are up to 200 nt in length. This length allows for filtration of RNA-seq data by discarding transcripts greater than the standard 200 nt. lncRNA may be classified as overlapping, intronic, exonic, and intron–exon based on location in the genome (Ma, Bajic, & Zhang, 2013; Rai, Alam, Lightfoot, Gurha, & Afzal, 2018). lncRNAs are formed from loci throughout the genome and in plants result from RNA pol IV and V coding followed by 5' capping, polyadenylation, and alternative splicing. Several techniques are used to distinguish them from mRNA through examining size, mass spectrometry data, and protein-coding capacity (Wang & Chekanova, 2017). With respect to function, plant lncRNAs are known to have control over neighboring transcription factors and ncRNA, structural regulation, and epigenetic changes (Karlik, Ari, & Gozukirmizi, 2019; Marchese, Raimondi, & Huarte, 2017).

To identify lncRNAs, library preparation involves rRNA depletion followed by poly-A enrichment and paired-end sequencing. It is necessary to map to a known reference genome to create transcript models or to perform de novo assembly in the situation that there is no reference genome available (Haas et al., 2013). It is possible to apply both approaches to create more accurate lncRNA models and construct the transcriptome. The transcript dataset is then classified into noncoding or coding based on examination of open reading frames alongside sequence conservation and is used to identify genomic loci capable of producing lncRNA. To classify transcripts, alignment is used to remove overlapping protein-coding genes based on reference genomes (NCBI, Ensembl) or querying protein databases for shared sequences (BLAST). An alignment-free method is then applied by examining the sequence for coding potential based on open reading frames to avoid incorrectly classifying sequences (Budak, Kaya, & Cagirici, 2020; Ilott & Ponting, 2013).

Next-generation sequencing has dramatically improved in recent years and these advancements allow for effective identification of different splicing isoforms and junction sequences. This is followed by structural and expression analysis using Cap analysis gene expression, RNA-seq, ChIP-seq with RNA pol II, V, and V to show RNA pol binding and transcription of lncRNA, detection of histone marks, transcriptional analysis, and other techniques to validate the lncRNA. Genome-wide approaches such as CAGE/TSS-seq, GRO-seq, and PAS-seq may also be applied to better map

TABLE 11.2 Collection of phasiRNA, tasiRNA, and NAT-siRNA predictive tools and relevant databases.

Annotation tools				
PhasiRNA				
Tool	Language	Platform	Description	URL
PhaseTank (2015) v1.0	Perl	Linux	PhaseTank may be applied to systemically characterize phasiRNAs/tasiRNAs and their regulatory cascades. The program offers one command analysis and thus provides high ease-of-use.	http://phasetank.sourceforge.net/
phasiRNAClassifier (2018) v1	Python	Linux	phasiRNAClassifier is a machine learning repository which provides scripts for generating sequence and structural features as well as classifying phasiRNA.	https://github.com/pupatel/phasiRNAClassifier
findPhasiRNAs (2019)	Python, R	Linux	This pipeline supports the identification of genomic loci where there is a strong indication of phasing and outputs <i>P</i> -values, phasing scores, and phasing structure.	https://github.com/Wiselab2/findPhasiRNAs
<i>tasiRNAs</i>				
TasExpAnalysis (2014)	–	–	This is a tool present as part of the tasiRNAdb which was developed to map small RNA and degradome libraries to a TAS, to perform phasing analysis, and to analyze TAS cleavage.	http://bioinfo.jit.edu.cn/tasiRNADatabase/
pssRNAMiner (2008)	–	–	pssRNAMiner maps input sRNAs against transcript/genomic sequences and identifies phased sRNA clusters.	http://bioinfo3.noble.org/pssRNAMiner/
<i>NAT-siRNAs</i>				
NATpipe (2016)	Perl	Linux	NATpipe is a pipeline used for discovery of NATs from de novo assembled transcriptomes using sRNA sequencing data. It allows users to search for phasiRNAs within NAT pairs.	http://www.bioinfolab.cn/NATpipe/NATpipe.zip
NATpare (2020)	Java, Javascript	Any platform	NATpare is a pipeline used for prediction and functional analysis of nat-siRNA. It uses multiple plant species as benchmarks and benefits from low resource requirements.	https://github.com/sRNAworkbench/UEA_sRNA_Workbench/
NASTI-seq (2020)	R	Linux, Windows	A pipeline for Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data.	https://ohlerlab.mdc-berlin.de/software/NASTIseq_104/
<i>Databases</i>				
<i>tasiRNAs</i>				
Database	Description			URL
tasiRNAdb (2014)	Holds information on TAS loci, tasiRNA sequences, tasiRNA targets and the like spanning a wide range of pathways and plants.			http://bioinfo.jit.edu.cn/tasiRNADatabase/
<i>tsRNA</i>				
PtRFdb (2018)	PtRFdb stores tRFs from 10 plant species along with information on tRF type, sequence, and genomic coordinates.			http://www.nipgr.ac.in/PtRFdb/
tRex (2018)	A web portal which provides access to ~1.4 million tRFs in <i>Arabidopsis thaliana</i> developed from tRNA annotations and published sRNA-seq datasets. Offers various search functionalities.			http://combio.pl/trex

transcript boundaries (Kashi, Henderson, Bonetti, & Carninci, 1859; Wang et al., 2013). To categorize lncRNA into overlapping, intronic, exonic, and intron–exon groups, it is necessary to examine their position and orientation. lncRNAs are genic (intron–exon, exon, intron), intergenic (between protein-coding genes), bidirectional (transcribed from a protein-coding gene promoter in an opposite direction), enhancer associated, or promoter associated. These are then further broken down into sense or antisense (Ma et al., 2013; Rai et al., 2018; St-Laurent, Wahlestedt, & Kapranov, 2015). Several barriers in classification exist, including lncRNAs may be shorter than 200 nt leading to false negatives, lncRNA candidates may encode proteins, RNA pol II offers low fidelity (indicating a need to identify transcripts from RNA pol IV or V), the lack of conserved motifs, tissue specificity, the presence of both spliced and unspliced RNA, and few to no validated predictive algorithms. Further, lncRNA lacks a uniform naming or annotation system and this leads to challenges in curating databases. Several predictive tools do exist; however, there are few plant-specific tools (Iwakiri, Hamada, & Asai, 1859). Examples of such tools are described in Table 11.3.

TABLE 11.3 Most common tools for the annotation of lncRNAs in plants.

Annotation tools				
Tool	Language	Platform	Description	URL
FEELnc (2017) v.0.1.1	Perl	Linux	An alignment-free program that annotates lncRNAs based on a Random Forest model trained with multi- <i>k</i> -mer frequencies and relaxed open reading frames.	https://github.com/tderrien/FEELnc
PLncPRO (2017) v1.2.2	Python	Linux	PLncPRO is used for the prediction of lncRNAs in plants using transcriptome data. PLncPRO is based on machine learning and uses random forest algorithm. This tool has high prediction accuracy and is designed for plants.	http://ccbb.jnu.ac.in/plncpro/
Evolinc-I/ Evolinc-II (2017) v1.7.5	Shell, Python, R	Linux	Evolinc is a pipeline designed to facilitate long intergenic noncoding RNA (lincRNA) discovery. It has two modules: the first (Evolinc-I) is an identification workflow that facilitates expression analysis and visualization of lincRNA. The second (Evolinc-II) offers transcriptomic and genomic comparative analysis to find phylogenetic depth of lincRNA locus conservation among related species.	https://github.com/Evolinc
Annocript (2015) v2.0.1	Perl	Linux	Annocript is a pipeline used to annotate de novo generated transcriptomes. It executes blast analysis with UniProt, NCBI Conserved Domain Database, and Nucleotide division and also provides annotations from Gene Ontology, the Enzyme Commission, and UniPathways.	https://github.com/frankMusacchia/Annocript
lncRNA-screen (2017)	Shell, R	Linux	lncRNA-screen is a pipeline for screening lncRNA transcripts over large multimodal datasets. It provides an automated pipeline which performs RNA-seq alignment, assembly, quality assessment, transcript filtration, novel lncRNA identification, expression level quantification, and more.	https://github.com/NYU-BFX/lncRNA-screen
Databases				
Tool	Description			URL
PLncDB (2013)	PLncDB consists of ~16k <i>Arabidopsis thaliana</i> lncRNAs covering genomic information, expression levels, genome browser visualization, and the like.			http://chualab.rockefeller.edu/gbrowse2/homepage.html
lncRNAdb (2011) v2.0	lncRNAdb provides users with a comprehensive reference of plant and other eukaryotic lncRNA. Includes sequence information, expression profiles, and associated literature.			http://lncrnadb.org/
CANTATAdb (2016) v2.0	Stores lncRNA from RNA-seq data of 39 plant species, covering expression information, coding potential, and other lncRNA loci information.			http://cantata.amu.edu.pl/

11.4 Circular RNA

Circular RNA (circRNA) is a class of ncRNA produced from transcription of protein-coding genes in a novel “backsplicing” mechanism in which a downstream splice donor site is covalently linked to an upstream acceptor site which creates a circular form. The product circRNA may be composed of one or more exons or introns. A wide array of circRNA has been detected in the literature in both humans and plants, though they possess relatively low, variable expression levels which previously led researchers to believe that they were simply by-products. It is now known that they may possess regulatory potential and may operate at the transcriptional and posttranscriptional level by regulating parent gene expression and ability to cause loss of function in miRNA. In addition, protein-coding circRNA has been found in humans, though little information on them exists in plants. Notably, circRNA is highly stable and this property may be the cause for their abundance in certain situations (Zhao, Chu, & Jiao, 2019; Ashwal-Fluss et al., 2014).

To identify circRNA with RNA-seq data, a search is performed to determine if reads have a reversed direction of transcription. This is performed by searching for chimeric or backspliced reads. Alternatively, it is possible to search for backspliced reads inclusive of all exon–intron circularizations and then performing experimental validation. Gene annotations are sparse for this particular class and as a result, searching for circRNA loci may lead to erroneous results. Algorithms rely on identifying sequences spanning the junction between the downstream and upstream sequences, after which these junctions spanning regions, using the previously described exon–intron combinations, are categorized by mapping to the site of genomic origin. The groups are exonic, intronic, and exon–intronic circRNA. Efforts to improve the process of identifying backsplicing junctions involve using reads of the loci in multiple samples and observing read counts (Yuan et al., 2018; Zhao et al., 2017).

TABLE 11.4 Annotation tools and databases for circRNA.

Annotation tools				
Tool	Language	Platform	Description	URL
find_circ2 (2013) v1	Python	Linux	A repository coded in python that may be used to detect backspliced sequencing reads, indicative of circular RNA (circRNA) in RNA-seq data.	https://github.com/rajewsky-lab/find_circ2
CircExplorer2 (2016) v2.3.8	Python	Linux	A program using single-read and paired-end reads designed to annotate circular RNA with high accuracy and low memory usage. It allows for circular RNA alignment tool application, de novo assembly of transcripts, characterization of backsplicing events, and other related functions.	https://github.com/YangLab/CIRCexplorer2
UROBORUS (2016) v2.0.0	Perl	Linux	A computational pipeline used to detect circRNAs in total RNA-seq data. It benefits from its ease of use and efficiency which can allow detection even when expression levels are low and without RNase R application to samples.	https://github.com/WGLab/UROBORUS
KNIFE (2018) v1.4	Python, Shell, Perl	Linux	An algorithm which improves the sensitivity and specificity of circRNA detection from RNA-seq data.	https://github.com/lindaszabo/KNIFE
CIRCfinder	Python	Linux	A pipeline to map junction reads for circRNAs. This pipeline allows for determination of the boundaries of circRNA to facilitate downstream analysis.	https://github.com/YangLab/CIRCfinder
Databases				
Database	Description			URL
PlantcircBase (2017) v5.0	Contains RNA-seq datasets derived from 19 plant species and provides predictions, information, and literature. It provides search functionalities for sequence and structural visualization.			http://ibi.zju.edu.cn/plantcircbase/
AtCircDB (2019) v2.0	A database on tissue-specific circRNAs of <i>Arabidopsis thaliana</i> with annotations and a visualization platform.			http://deepbiology.cn/circRNA/
CircFunBase (2019)	Documents circRNAs of 7 plant species and other animals with information on function and interaction networks for both validated and predicted circRNA.			http://bis.zju.edu.cn/CircFunBase

As with many of the RNA discussed in this chapter, few tools exist to study this novel class, particularly in plants. Those which do exist vary with respect to detection, categorization, and the like. However, issues may arise with computational tools given that it is difficult to distinguish between circRNA and linear RNA based purely on the detection of a backsplicing junction. Further, trans-splicing events, artifacts due to alignment errors, experimental errors, genomic rearrangements, exon repeats, homologous exons, and the like tend to complicate identification. Given these factors, it is suggested to use multiple algorithms together. To prepare such samples, several approaches are possible including selecting libraries with/without poly(A) tails. One common approach applies RNase R to remove linear RNA from a sample followed by validation with paired-end reads (Zhang, Li, & Chen, 2020; Tang, Hao, Zhu, Zhang, & Li, 2018).

With respect to tools available for the study of circRNA, there are few sound examples. PcircRNA_finder is one of the few algorithms capable of predicting plant-specific circRNA (Chen et al., 2016). Databases include, but are not limited to, AtCircDB, PlantcircBase, PlantCircNet, and CropCircDB (Ye, Wang, & Li, 2019; Wang, Wang, & Guo, 2019). These databases include information on circRNA loci, backsplicing read junctions, read counts, and expression data. However, nomenclature, categorization by uniform identifiers, limited data on circRNA formed by alternative splicing, and interoperability regarding circRNA data (ability to integrate with existing databases) remains a concern. Annotation tools and databases for circRNA may be found in Table 11.4.

11.5 Chimeric RNA

Gene fusions are hallmarks of many cancer types and have unique cytogenetic signatures which allow for them to serve as effective biomarkers. The most well-characterized fusion in humans was the chromosomal abnormality, the Philadelphia chromosome, in which BCR and ABL1 genes combine to create the BCR-ABL1 fusion which encodes a kinase and critical biomarker of Chronic Myeloid Leukemia. Since that discovery, gene fusions have been found in both cancerous and normal cell types, indicating a potentially significant contributor to the functional genome. Chimeric transcripts may act as lncRNA or encode chimeric proteins. Such fusions are formed from fusion transcripts, or chimeric RNA. That is, a hybrid RNA composed of transcripts from two different genes. The mechanisms of such chimerization are being studied but may be attributed to changes in the genome as with the Philadelphia Chromosome as well as noncanonical methods, including trans-splicing of precursor mRNA or cis-splicing of adjacent genes (cis-SAGE). Each of these splicing events is mediated by spliceosome complexes (Singh, Qin, & Kumar, 2020).

TABLE 11.5 Tools and databases for the annotation of chimeric transcripts.

Annotation tools				
Tool	Language	Platform	Description	URL
Chimeraviz (2017)	R	–	A Bioconductor package in R that automates the creation of chimeric RNA visualizations. It supports input from 9 other fusion-finder tools.	https://www.bioconductor.org/packages/release/bioc/html/chimeraviz.html
EricScript (2012)	Python	Linux	EricScript is a computational framework used for the discovery of gene fusions in paired-end RNA-seq data.	https://sites.google.com/site/bioericscript/
JAFFA (2015)	–	Linux	A sensitive fusion detection tool which performs extremely well when applied to reads greater than 100 bp. It compares cancer transcriptomes to reference transcriptomes and allows for inference of the cancer transcriptome using reads and de novo assembly.	https://github.com/Oshlack/JAFFA
Databases				
Tool	Description			URL
ChiTaRS 5.0 (2019)	A comprehensive chimeric transcript repository, with ~11,000-annotated entries from 8 species.			http://chitars.md.biu.ac.il/

Fusion transcript formation has been studied in a number of model organisms, including humans, zebrafish, and plants (Zhang, Guo, & Hu, 2010; Koller, Fromm, Galun, & Edelman, 1987; Kawasaki et al., 1999). The advent and advancement of high-throughput sequencing technologies leads to the identification of fusion transcripts in RNA-Seq datasets. Several tools exist to better characterize and study this class of RNA. Fusion transcript prediction tools using RNA-seq data to generate potential fusion candidates include EricScript, SOAPfuse, and JAFFA (Benelli et al., 2012; Jia, Qiu, & He, 2013; Davidson, Majewski, & Oshlack, 2015). Several fusion transcript databases exist, including ChiTaRS, FusionCancer, ChimerDB, FusionHub, and AtFusionDb which span the range of species such as humans, mice, flies, and Arabidopsis (Balamurali et al., 2019; Singh, Zahra, Das, & Kumar, 2019; Wang, Wu, Liu, Wu, & Dong, 2015; Jang, Jang, & Kim, 2019; Panigrahi, Jere, & Anamika, 2018). Broadly, further studies must be performed to better characterize and predict chimeric RNA in plants. This is in part due to the previous widely held belief that such fusions are limited to cancerous human tissue. A selection of existing tools and databases may be found in Table 11.5.

References

- Alves, C. S., Vicentini, R., Duarte, G. T., Pinoti, V. F., Vincentz, M., & Nogueira, F. T. S. (2017). Genome-wide identification and characterization of tRNA-derived RNA fragments in land plants. *Plant Molecular Biology*, 93, 35–48. Available from <https://doi.org/10.1007/s11103-016-0545-9>.
- Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., . . . Kadener, S. (2014). circRNA biogenesis competes with pre-mRNA splicing. *Molecular Cell*, 56, 55–66. Available from <https://doi.org/10.1016/j.molcel.2014.08.019>.
- Axtell, M. (2013). ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA (New York, N.Y.)*, 19(6), 740–751. Available from <https://doi.org/10.1261/rna.035279.112>.
- Axtell, M. J. (2013). Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology*, 64, 137–159. Available from <https://doi.org/10.1146/annurev-arplant-050312-120043>.
- Bailey-Serres, J., Zhai, J., & Seki, M. (2020). The dynamic kaleidoscope of RNA biology in plants. *Plant Physiology*, 182(1), 1–9. Available from <https://doi.org/10.1104/pp.19.01558>.
- Balamurali, D., Gorohovski, A., Detroja, R., Palande, V., Raviv-Shay, D., & Frenkel-Morgenstern, M. (2019). ChiTaRS 5.0: The comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps. *Nucleic Acids Research*. Available from <https://doi.org/10.1093/nar/gkz1025>.
- Benelli, M., Pescucci, C., Marseglia, G., Severgnini, M., Torricelli, F., & Magi, A. (2012). Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics (Oxford, England)*, 28(24), 3232–3239. Available from <https://doi.org/10.1093/bioinformatics/bts617>.
- Berezikov, E., Cuppen, E., & Plasterk, R. H. A. (2006). Approaches to microRNA discovery. *Nature Genetics*, 38, S2. Available from <https://doi.org/10.1038/ng1794>.
- Borges, F., & Martienssen, R. (2015). The expanding world of small RNAs in plants. *Nature Reviews. Molecular Cell Biology*, 16(12), 727–741. Available from <https://doi.org/10.1038/nrm4085>.
- Bortolomeazzi, M., Gaffo, E., & Bortoluzzi, S. (2017). A survey of software tools for microRNA discovery and characterization using RNA-seq. *Briefings in Bioinformatics*, 20, 918–930. Available from <https://doi.org/10.1093/bib/bbx148>.
- Budak, H., Kaya, S. B., & Cagirici, H. B. (2020). Long non-coding RNA in plants in the era of reference sequences. *Frontiers of Plant Science*, 11, 276.
- Cao, H., Wahlestedt, C., & Kapranov, P. (2018). Strategies to annotate and characterize long noncoding RNAs: Advantages and pitfalls. *Trends in Genetics*, 34, 704–721. Available from <https://doi.org/10.1016/j.tig.2018.06.002>.
- Carbonell, A. (2019). Secondary small interfering RNA-based silencing tools in plants: An update. *Frontiers of Plant Science*, 10. Available from <https://doi.org/10.3389/fpls.2019.00687>.
- Chen, L., Yu, Y., Zhang, X., Liu, C., Ye, C., & Fan, L. (2016). PcircRNA_finder: A software for circRNA prediction in plants. *Bioinformatics (Oxford, England)*, 32(22), 3528–3529. Available from <https://doi.org/10.1093/bioinformatics/btw496>.
- Chen, C., Feng, J., Liu, B., Li, J., Feng, L., Yu, X., . . . Xia, R. (2019). sRNAanno—A database repository of uniformly-annotated small RNAs in plants. *bioRxiv*, 771121. Available from <https://doi.org/10.1101/771121>.
- Chen, H. M., Li, Y. H., & Wu, S. H. (2007). Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 3318–3323. Available from <https://doi.org/10.1073/pnas.0611119104>.
- Davidson, N. M., Majewski, I. J., & Oshlack, A. (2015). JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Medicine*, 7, 43. Available from <https://doi.org/10.1186/s13073-015-0167-x>.
- Deng, P., Muhammad, S., Cao, M., & Wu, L. (2018). Biogenesis and regulatory hierarchy of phased small interfering RNAs in plants. *Plant Biotechnology Journal*, 16(5), 965–975. Available from <https://doi.org/10.1111/pbi.12882>.
- Fei, Q., Xia, R., & Meyers, B. C. (2013). Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant Cell*, 25(7), 2400–2415. Available from <https://doi.org/10.1105/tpc.113.114652>.
- Feng, L., Zhang, F., Zhang, H., Zhao, Y., Meyers, B. C., & Zhai, J. (2020). An online database for exploring over 2,000 Arabidopsis small RNA Libraries. *Plant Physiology*, 182, 685–691. Available from <https://doi.org/10.1104/pp.19.00959>.

- Gomes, C. P. C., Cho, J. H., Hood, L., Franco, O. L., Pereira, R. W., & Wang, K. (2013). A review of computational tools in microRNA discovery. *Frontiers in Genetics*, 4. Available from <https://doi.org/10.3389/fgene.2013.00081>.
- Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34, D140–D144. Available from <https://doi.org/10.1093/nar/gkj112>.
- Guo, Q., Liu, Q., Smith, N. A., Liang, G., & Wang, M. B. (2016). RNA Silencing in plants: Mechanisms, technologies and applications in horticultural crops. *Current Genomics*, 17(6), 476–489. Available from <https://doi.org/10.2174/1389202917666160520103117>.
- Guo, Q., Qu, X., & Jin, W. (2015). PhaseTank: Genome-wide computational identification of phasiRNAs and their regulatory cascades. *Bioinformatics (Oxford, England)*, 31, 284–286. Available from <https://doi.org/10.1093/bioinformatics/btu628>.
- Guo, Z., Kuang, Z., Wang, Y., Zhao, Y., Tao, Y., Cheng, C., Yang, J., Lu, X., Hao, C., Wang, T., et al. (2019). PmiREN: A comprehensive encyclopedia of plant miRNAs. *Nucleic Acids Research*. Available from <https://doi.org/10.1093/nar/gkz894>.
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8, 1494–1512. Available from <https://doi.org/10.1038/nprot.2013.084>.
- Huang, C., Wang, H., Hu, P., Hamby, R., & Jin, H. (2019). Small RNAs – big players in plant-microbe interactions. *Cell Host & Microbe*, 26(2), 173–182. Available from <https://doi.org/10.1016/j.chom.2019.07.021>.
- Iltot, N. E., & Ponting, C. P. (2013). Predicting long non-coding RNAs using RNA sequencing. *Methods (San Diego, Calif.)*, 63, 50–59. Available from <https://doi.org/10.1016/j.ymeth.2013.03.019>.
- Iwakiri, J., Hamada, M., & Asai, K. (1859). Bioinformatics tools for lncRNA research. *Biochimica et Biophysica Acta – Gene Regulatory Mechanisms*, 2016, 23–30.
- Jang, Y., Jang, I., Kim, S., et al. (2019). ChimerDB 4.0: An updated and expanded database of fusion genes. *Nucleic Acids Research*. Available from <https://doi.org/10.1093/nar/gkz1013>.
- Jia, W., Qiu, K., He, M., et al. (2013). SOAPfuse: An algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biology*, 14, R12. Available from <https://doi.org/10.1186/gb-2013-14-2-r12>.
- Johnson, C., Bowman, L., Adai, A. T., Vance, V., & Sundaresan, V. (2007). CSRDB: A small RNA integrated database and browser resource for cereals. *Nucleic Acids Research*, 35, D829–D833. Available from <https://doi.org/10.1093/nar/gkl991>.
- Jones-Rhoades, M. W., Bartel, D. P., & Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology*, 57, 19–53. Available from <https://doi.org/10.1146/annurev.arplant.57.032905.105218>.
- Kakrana, A., Li, P., Patel, P., Hammond, R., Anand, D., Mathioni, S., & Meyers, B. (2017). PHASIS: A computational suite for de novo discovery and characterization of phased, siRNA-generating loci and their miRNA triggers. *bioRxiv*, 158832. Available from <https://doi.org/10.1101/158832>.
- Kamthan, A., Chaudhuri, A., Kamthan, M., & Datta, A. (2015). Small RNAs in plants: Recent development and application for crop improvement. *Frontiers of Plant Science*, 06. Available from <https://doi.org/10.3389/fpls.2015.00208>.
- Karlik, E., Ari, S., & Gozukirmizi, N. (2019). LncRNAs: Genetic and epigenetic effects in plants. *Biotechnology & Biotechnological Equipment*, 33(1), 429–439. Available from <https://doi.org/10.1080/13102818.2019.1581085>.
- Kashi, K., Henderson, L., Bonetti, A., & Carninci, P. (1859). Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochimica et Biophysica Acta – Gene Regulatory Mechanisms*, 2016, 3–15. Available from <https://doi.org/10.1016/j.bbgrm.2015.10.010>.
- Kawasaki, T., Okumura, S., Kishimoto, N., Shimada, H., Higo, K., & Ichikawa, N. (1999). RNA maturation of the rice SPK gene may involve trans-splicing. *The Plant Journal*, 18(6), 625–632. Available from <https://doi.org/10.1046/j.1365-313x.1999.00493.x>.
- Koller, B., Fromm, H., Galun, E., & Edelman, M. (1987). Evidence for in vivo trans splicing of pre-mRNAs in tobacco chloroplasts. *Cell*, 48(1), 111–119. Available from [https://doi.org/10.1016/0092-8674\(87\)90361-8](https://doi.org/10.1016/0092-8674(87)90361-8).
- Komiya, R. (2017). Biogenesis of diverse plant phasiRNAs involves an miRNA-trigger and Dicer-processing. *Journal of Plant Research*, 130(1), 17–23. Available from <https://doi.org/10.1007/s10265-016-0878-0>.
- Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). MiRBase: From microRNA sequences to function. *Nucleic Acids Research*, 47, D155–D162. Available from <https://doi.org/10.1093/nar/gky1141>.
- Kozomara, A., & Griffiths-Jones, S. (2014). MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42, 68–73. Available from <https://doi.org/10.1093/nar/gkt1181>.
- Kumar, P., Kuscus, C., & Dutta, A. (2016). Biogenesis and function of transfer RNA-related fragments (tRFs). *Trends in Biochemical Sciences*, 41, 679–689. Available from <https://doi.org/10.1016/j.tibs.2016.05.004>.
- Kumar, P., Mudunuri, S. B., Anaya, J., & Dutta, A. (2015). tRFdb: A database for transfer RNA. *Nucleic Acids Research*, 43, D141–D145. Available from <https://doi.org/10.1093/nar/gku1138>.
- Lee, Y., Jeon, K., Lee, J. T., Kim, S., & Kim, V. N. (2002). MicroRNA maturation: Stepwise processing and subcellular localization. *The EMBO Journal*, 21, 4663–4670. Available from <https://doi.org/10.1093/emboj/cdf476>.
- Lei, J., & Sun, Y. (2014). miR-PREFeR: An accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics*, 30, 2837–2839. Available from <https://doi.org/10.1093/bioinformatics/btu380>.
- Leinonen, R., Sugawara, H., & Shumway, M. (2011). International nucleotide sequence database collaboration. The sequence read archive. *Nucleic Acids Research*, 39(Database issue), D19–D21. Available from <https://doi.org/10.1093/nar/gkq1019>.
- Li, S., Xu, Z., & Sheng, J. (2018). tRNA-derived small RNA: A novel regulatory small non-coding RNA. *Genes (Basel)*, 9.

- Liao, P., Li, S., Cui, X., & Zheng, Y. (2018). A comprehensive review of web-based resources of non-coding RNAs for plant science research. *International Journal of Biological Sciences*, *14*, 819–832. Available from <https://doi.org/10.7150/ijbs.24593>.
- Loss-Morais, G., Waterhouse, P. M., & Margis, R. (2013). Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. *Biology Direct*, *8*.
- Ma, L., Bajic, V. B., & Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biology*, *10*(6), 925–933. Available from <https://doi.org/10.4161/rna.24604>.
- Ma, X., Tang, Z., Qin, J., & Meng, Y. (2015). The use of high-throughput sequencing methods for plant microRNA research. *RNA Biology*, *12*, 709–719. Available from <https://doi.org/10.1080/15476286.2015.1053686>.
- Marchese, F., Raimondi, I., & Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. *Genome Biology*, *18*(1). Available from <https://doi.org/10.1186/s13059-017-1348-2>fragments. (2015). *Nucleic Acids Research*, *43*, D141–D145. doi:10.1093/nar/gku1138.
- Martinez, G., Choudury, S. G., & Slotkin, R. K. (2017). tRNA-derived small RNAs target transposable element transcripts. *Nucleic Acids Research*, *45*, 5142–5152. Available from <https://doi.org/10.1093/nar/gkx103>.
- Mendes, N. D., Freitas, A. T., & Sagot, M. F. (2009). Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Research*, *37*, 2419–2433. Available from <https://doi.org/10.1093/nar/gkp145>.
- Morgado, L., & Johannes, F. (2019). Computational tools for plant small RNA detection and categorization. *Briefings in Bioinformatics*, *20*, 1181–1192. Available from <https://doi.org/10.1093/bib/bbx136>.
- Neilsen, C. T., Goodall, G. J., & Bracken, C. P. (2012). IsomiRs – The overlooked repertoire in the dynamic microRNAome. *Trends in Genetics: TIG*, *28*, 544–549. Available from <https://doi.org/10.1016/J.TIG.2012.07.005>.
- Panigrahi, P., Jere, A., & Anamika, K. (2018). FusionHub: A unified web platform for annotation and visualization of gene fusion events in human cancer. *PLoS One*, *13*(5), e0196588. Available from <https://doi.org/10.1371/journal.pone.0196588>, Published 2018 May 1.
- Rai, M. I., Alam, M., Lightfoot, D. A., Gurha, P., & Afzal, A. J. (2018). Classification and experimental identification of plant long non-coding RNAs. *Genomics*. Available from <https://doi.org/10.1016/J.YGENO.2018.04.014>.
- Sablok, G., Srivastva, A. K., Suprasanna, P., Baev, V., & Ralph, P. J. (2015). isomiRs: Increasing evidences of isomiRs complexity in plant stress functional biology. *Frontiers of Plant Science*, *6*. Available from <https://doi.org/10.3389/fpls.2015.00949>.
- Shin, S. Y., & Shin, C. (2016). Regulatory non-coding RNAs in plants: Potential gene resources for the improvement of agricultural traits. *Plant Biotechnology Reports*, *10*, 35–47. Available from <https://doi.org/10.1007/s11816-016-0389-4>.
- Signal, B., Gloss, B. S., & Dinger, M. E. (2016). Computational approaches for functional prediction and characterisation of long noncoding RNAs. *Trends in Genetics: TIG*, *32*, 620–637. Available from <https://doi.org/10.1016/j.tig.2016.08.004>.
- Singh, A., Zahra, S., Das, D., & Kumar, S. (2019). AtFusionDB: A database of fusion transcripts in *Arabidopsis thaliana*. *Database*, 2019. Available from <https://doi.org/10.1093/database/bay135>.
- Singh, S., Qin, F., Kumar, S., et al. (2020). The landscape of chimeric RNAs in non-diseased tissues and cells. *Nucleic Acids Research*, *48*(4), 1764–1778. Available from <https://doi.org/10.1093/nar/gkz1223>.
- St-Laurent, G., Wahlestedt, C., & Kapranov, P. (2015). The landscape of long noncoding RNA classification. *Trends in Genetics: TIG*, *31*, 239–251. Available from <https://doi.org/10.1016/j.tig.2015.03.007>.
- Szakonyi, D., Confraria, A., Valerio, C., Duque, P., & Staiger, D. (2019). Editorial: Plant RNA biology. *Frontiers of Plant Science*, *10*. Available from <https://doi.org/10.3389/fpls.2019.00887>.
- Tang, B., Hao, Z., Zhu, Y., Zhang, H., & Li, G. (2018). Genome-wide identification and functional analysis of circRNAs in *Zea mays*. *PLoS One*, *13*, e0202375. Available from <https://doi.org/10.1371/journal.pone.0202375>.
- Thody, J., Folkles, L., & Moulton, V. (2020). NATpare: A pipeline for high-throughput prediction and functional analysis of nat-siRNAs. *Nucleic Acids Research*, *48*(12), 6481–6490. Available from <https://doi.org/10.1093/nar/gkaa448>.
- Thompson, A., Zielezinski, A., Plewka, P., Szymanski, M., Nuc, P., Szweykowska-Kulinska, Z., ... Karlowski, W. M. (2018). tRex: A web portal for exploration of tRNA-derived fragments in *Arabidopsis thaliana*. *Plant & Cell Physiology*, *59*, e1. Available from <https://doi.org/10.1093/pcp/pcx173>.
- Vivek, A., Zahra, S., & Kumar, S. (2019). From current knowledge to best practice: A primer on viral diagnostics using deep sequencing of virus-derived small interfering RNAs (vsiRNAs) in infected plants. *Methods (San Diego, California)*. Available from <https://doi.org/10.1016/j.ymeth.2019.10.009>.
- Wang, H. V., & Chekanova, J. A. (2017). Long noncoding RNAs in plants. *Advances in Experimental Medicine and Biology*, *1008*, 133–154. Available from https://doi.org/10.1007/978-981-10-5203-3_5.
- Wang, J., Mei, J., & Ren, G. (2019). Plant microRNAs: Biogenesis, homeostasis, and degradation. *Frontiers of Plant Science*, *10*. Available from <https://doi.org/10.3389/fpls.2019.00360>.
- Wang, K., Wang, C., Guo, B., et al. (2019). CropCircDB: A comprehensive circular RNA resource for crops in response to abiotic stress. *Database*, 2019. Available from <https://doi.org/10.1093/database/baz053>.
- Wang, Y., Wu, N., Liu, J., Wu, Z., & Dong, D. (2015). FusionCancer: A database of cancer fusion genes derived from RNA-seq data. *Diagnostic Pathology*, *10*, 131. Available from <https://doi.org/10.1186/s13000-015-0310-4>, Published 2015 Jul 28.
- Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., & Li, W. (2013). CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research*, *41*, 1–7. Available from <https://doi.org/10.1093/nar/gkt006>.
- Wu, F., Chen, Y., Tian, X., et al. (2017). Genome-wide identification and characterization of phased small interfering RNA genes in response to *Botrytis cinerea* infection in *Solanum lycopersicum*. *Scientific Reports*, *7*, 3019. Available from <https://doi.org/10.1038/s41598-017-02233-x>.

- Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., & Qi, Y. (2010). DNA methylation mediated by a microRNA pathway. *Molecular Cell*, 38, 465–475. Available from <https://doi.org/10.1016/j.molcel.2010.03.008>.
- Xie, Y., Yao, L., Yu, X., Ruan, Y., Li, Z., & Guo, J. (2020). Action mechanisms and research methods of tRNA-derived small RNAs. *Signal Transduction and Targeted Therapy*, 5(1). Available from <https://doi.org/10.1038/s41392-020-00217-4>.
- Ye, J., Wang, L., Li, S., et al. (2019). AtCircDB: A tissue-specific database for Arabidopsis circular RNAs. *Briefings in Bioinformatics*, 20(1), 58–65. Available from <https://doi.org/10.1093/bib/bbx089>.
- Yu, D., Meng, Y., Zuo, Z., Xue, J., & Wang, H. (2016). NATpipe: An integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes. *Scientific Reports*, 6(1). Available from <https://doi.org/10.1038/srep21666>.
- Yuan, Y., Cai, Y., Xiang, L., Cai, C., Cheng, J., Wang, L., Wu, C., Shi, Y., Luo, J., He, L., et al. (2018). Identification of circularRNAs and their targets in *Gossypium* under *Verticillium wilt* stress based on RNA-seq. *PeerJ*, 6, e4500. Available from <https://doi.org/10.7717/peerj.4500>.
- Zhang, C., Li, G., Zhu, S., Zhang, S., & Fang, J. (2013). tasiRNAdb: A database of ta-siRNA regulatory pathways. *Bioinformatics (Oxford, England)*, 30(7), 1045–1046. Available from <https://doi.org/10.1093/bioinformatics/bt746>.
- Zhang, G., Guo, G., Hu, X., et al. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research*, 20(5), 646–654. Available from <https://doi.org/10.1101/gr.100677.109>.
- Zhang, P., Li, S., & Chen, M. (2020). Characterization and function of circular RNAs in plants. *Frontiers in Molecular Biosciences*, 7. Available from <https://doi.org/10.3389/fmolb.2020.00091>.
- Zhang, X., Xia, J., Lii, Y. E., et al. (2012). Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. *Genome Biology*, 13(3), R20. Available from <https://doi.org/10.1186/gb-2012-13-3-r20>.
- Zhao, W., Chu, S., & Jiao, Y. (2019). Present scenario of circular RNAs (circRNAs) in plants. *Frontiers of Plant Science*, 10. Available from <https://doi.org/10.3389/fpls.2019.00379>.
- Zhao, W., Cheng, Y., Zhang, C., You, Q., Shen, X., Guo, W., & Jiao, Y. (2017). Genome-wide identification and characterization of circular RNAs by high throughput sequencing in soybean. *Scientific Reports*, 7, 5636. Available from <https://doi.org/10.1038/s41598-017-05922-9>.
- Zheng, Y., Wang, S., & Sunkar, R. (2014). Genome-wide discovery and analysis of phased small interfering RNAs in Chinese sacred lotus. *PLoS One*, 9, e113790. Available from <https://doi.org/10.1371/journal.pone.0113790>.
- Zheng, Y., Wang, Y., Wu, J., Ding, B., & Fei, Z. (2015). A dynamic evolutionary and functional landscape of plant phased small interfering RNAs. *BMC Biology*, 13, 32. Available from <https://doi.org/10.1186/s12915-015-0142-4>.
- Zhu, L., Ge, J., Li, T., Shen, Y., & Guo, J. (2019). tRNA-derived fragments and tRNA halves: The new players in cancers. *Cancer Letters*, 452, 31–37.

This page intentionally left blank

Molecular evolution, three-dimensional structural characteristics, mechanism of action, and functions of plant beta-galactosidases

Md. Anowar Hossain

Department of Biochemistry and Molecular Biology, University of Rajshahi, Rajshahi, Bangladesh

12.1 Introduction

Beta-galactosidase (BGAL, EC 3.2.1.23) is one of the oldest ubiquitous enzymes, which catalyzes the hydrolysis of nonreducing β -D-galactosyl residues from β -D-galactoside polymers. It is also called by various names such as lactase, β -lactosidase, maxilact, hydrolact, β -D-lactosidase, lactozym, trilactase, β -D-galactanase, oryzatym, sumiklat, β -D-galactanase, β -galase, *exo*- β -(1 \rightarrow 4)-D-galactanase, and *exo*- β -(1 \rightarrow 3)-D-galactanase based on their substrate-catalyzed reactions, sources, and mechanism of action (Dwevedi & Kayastha, 2010). It has been reported that BGAL has the ability to hydrolyze the terminal galactosyl residues from carbohydrates, glycoproteins, and galactolipids (Ross, Wagrzyn, MacRae, & Redgwell, 1994; Smith, Abbott, & Gross, 2002; Smith & Gross, 2000). BGALs are widely distributed in lower to higher organisms including bacteria, fungi, yeasts, plants, and animals (Husain, 2010). These enzymes have been classified into glycoside hydrolase (GH) families as GH1, GH2, GH35, and GH42 in CAZy (carbohydrate-active enzymes, <http://www.cazy.org/>) database. BGALs present in microorganisms mostly belong to GH1, GH2, and GH42. On the other hand, BGALs belonging to GH35 are found in a wide range of organisms including bacteria, fungi, animals, and plants (Dwevedi & Kayastha, 2010). On the basis of their substrate specificities, the plant BGALs (pBGALs) can be divided into two types: one that consists of *exo*- β -(1 \rightarrow 4)-galactanases particularly function on pectic β -(1 \rightarrow 4)-D galactan, and the other that acts on β -(1 \rightarrow 3) and β -(1 \rightarrow 6)-galactosyl linkages of arabinogalactan proteins but does not have hydrolytic activity against β -(1 \rightarrow 4)-galactan (Kotake et al., 2005).

BGALs have various physiological roles in plants including cell-wall expansion and degradation, and turnover of signaling molecules during ripening (Buckeridge & Reid, 1994; de Alcantara, Martim, Silva, Dietrich, & Buckeridge, 2006; Ross, Redgwell, & MacRae, 1993). Recently, the pBGALs have gained much interest for their involvement in the developmental stages and pectin degradation during fruit ripening in various plants including tomato (Carey et al., 1995; Moctezuma, Smith, & Gross, 2003; Pressey, 1983), muskmelon (Ranwala, Suematsu, & Masuda, 1992), kiwifruit (Ross et al., 1993), mango (Ali, Armugam, & Lazan, 1995), peach (Lee, Kang, Suh, & Byun, 2003), radish (Kotake et al., 2005), papaya (Lazan, Ng, Goh, & Ali, 2004), and apple (Yang et al., 2018). β -Galactosidase activity significantly increased in tomato fruits during ripening that suggested their roles in the breakdown of β (1 \rightarrow 4)-galactan side chains of pectin as part of the ripening process (Carey et al., 1995). Subsequently, it has been reported that downregulation of a ripening-related BGAL mRNA decreased the enzyme activity and freed galactose content and significantly retained the fruit firmness (Smith et al., 2002). Another report on pectin changes and pectin-modifying enzymes in Jonagold apples during postharvest softening showed that the BGAL was the key player for softening during ripening (Gwanpua et al., 2014). Our previous study showed that mango ripening-related enzymes such as BGAL, α -mannosidase, and beta-hexosaminidase changed significantly during the postharvest storage at different temperatures (Hossain, Rana, Kimura, & Roslan, 2014). Recently, Yang's group reported that BGAL activity and expression levels

of BGAL genes (Md β -Gal1, Md β -Gal2, and Md β -Gal5) significantly increased in “Fuji” and “Qinguan” apples during all stages of fruit developmental and were much higher in the mature fruits; indicating that pectin was degraded by BGALs (Yang et al., 2018).

The GH35 like other families contains multiple copies of BGAL genes in different plant species. At least 17 BGAL genes were reported from tomatoes, of which 6 were expressed during fruit development stages and ripening (Smith & Gross, 2000; Chandrasekar & Hoorn, 2016). In *Arabidopsis*, 17 putative BGAL genes were found to be expressed that were further divided into seven subgroups based on their sequence similarities. Subgroup-III included seven members that involved in the modification of pectic polysaccharides of cell-wall matrices (Ahn et al., 2007). Meanwhile, 15 BGAL genes were identified in rice, 1 of which encoded for a protein similar to animal BGAL and the rest 14 were grouped into a plant-specific subfamily of BGALs and few BGAL genes were located on the different chromosomes by segments duplication (Tanthanuch, Chantarangsee, Maneesan, & Ketudat-Cairns, 2008). Rice BGAL enzymes might play important roles in cell-wall polysaccharides, glycoproteins, and glycolipids metabolism. At least two BGAL isoforms were identified and characterized from the *Coffea arabica* genome (Figueiredo, Lashermes, & Araga, 2011). Recently, a comprehensive genome-wide analysis of *Brassica campestris* ssp. *chinensis* identified 16 BGAL genes (Liu, Gao, Lv, & Cao, 2013). Based on their conserved motifs, *Brassica* BGALs (BcBGALs) were classified into four groups and 7 out of the 16 BcBGAL genes had two copies, whereas one BcBGAL gene contained five copies. Exon-intron structures of different BcBGAL genes within the same group were very similar (Liu et al., 2013). Altogether the results obtained from the above observation, it is postulated that pBGALs under GH35 family have multiple copies of gene that might be generated through segmental gene duplications.

The determination of three-dimensional (3D) structure of an enzyme is a prerequisite to get a better understanding of the functional mechanism of an enzyme. Numerous X-ray solved crystal structures of BGALs belonging to GH35 family have been deposited to protein data bank (<https://www.rcsb.org/>). The first 3D structure of β -galactosidase from *Escherichia coli* (EcBGAL) was published in the Nature in 1994 (Jacobson, Zhang, Dubose, & Matthews, 1994). EcBGAL is a tetrameric structure of four identical polypeptide chains with a calculated molecular mass of 465 kDa. Each subunit contains five domains: jelly-roll type barrel (Domain 1), fibronectin type III-like barrels (Domain 2 and 4), β -sandwich (Domain 5), and the TIM (triose-phosphate isomerase)-type barrel (Domain 3). Central domain 3 houses the active site amino-acid residues. Similar to EcBGAL, crystal structure of *Penicillium* sp. BGAL has five distinct domains but the first domain is distorted TIM barrel that contains the catalytic site (Rojas et al., 2004). On the other hand, human BGAL consists of catalytic TIM-barrel domain, β -domain 1, and β -domain 2 (Ohto et al., 2012). The first and only X-ray crystal structure of pBGAL (TBG4) was solved at 1.65 Å resolution (pdb id: 3w5g) from tomato fruit by a Japanese research group (Eda, Ishimaru, & Tada, 2015; Masahiro Eda, Matsumoto, Ishimaru, & Tada, 2016). Recently, the phylogenetic relationship, homology modeling, docking, and mechanism of action of *Mangifera indica* BGAL (MiBGAL) have been elucidated (Hossain, Roslan, Karim, & Kimura, 2016). This chapter summarizes the molecular evolution, structural features, mechanism of action, and physiological functions of pBGALs.

12.2 Protein sequence features of plant beta-galactosidases

Numerous BGALs have been characterized based on the number of amino acids that resided in the polypeptide chain of active enzymes. The number of amino acids in the BGAL enzymes varies from higher organism to lower one. The smallest BGAL was found in bacteria (586–613 aa). The largest BGALs were found in fungi that contain 1002–1023 aa followed by plants (715–857 aa) and animals (647–677 aa) (Table 12.1) (Hossain et al., 2016). NCBI CD (conserved domain)-search tool (CDD V3.0–44354 PSSMs) was used to identify the CDs in the 67 BGAL protein sequences. All BGALs usually consist of GH35, GH42, LacA domain, and a BGAL multidomain, called “PLN03059” (Hossain et al., 2016). The pBGAL sequence possesses an additional unknown functional domain “DUF4185” including 244–324 aa and a galactose binding lectin domain with 750–827 aa (Fig. 12.1). Moreover, the pBGALs also contain a unique galactose binding lectin domain in the C-terminal region if they have more than 750 amino acids. Bacterial and animal BGALs possess the following common domains such as GH35, GH42, PLN03059, BGal-dom4.5, and Gal-lactin but some bacteria don’t have additional BGal-dom4.5 (Hossain et al., 2016). The functional roles of these additional domains are not yet clear. However, it was suggested that the Gal-lectin domain could play a role in substrate specificity of BGAL (Chandrasekar & Hoorn, 2016). Meanwhile, MiBGAL contained all types of domains in a complete multi-domain architecture (PLN03059) (Masahiro Eda et al., 2016). These domains were termed domains I, II, III, and IV due to their common presence in other proteins (e.g., domain-I-TIM-barrel domain). They receive different names when they are also present in different protein families. As can be seen, what is called a GH42 domain is part of the GH35 domain (and they are both parts of a TIM-barrel domain) (Hossain et al., 2016).

TABLE 12.1 The features of beta-galactosidase sequences used for phylogenetic analysis (Hossain et al., 2016).

Sl no.	GI number	Name used in the phylogenetic tree	Organisms	Taxonomy	No. of amino acids	Domains identified by CD-search	Signal peptide cleavage site	N-glycosylation sites
1.	1857333	Arthrobacter-BGAL	<i>Arthrobacter</i> sp.	Prokaryota (bacteria)	471	GH42, 35, PLN03059, LacA	No	No
2.	76097478	X_campestris-GalD	<i>Xanthomonas campestris</i>	Prokaryota (bacteria)	579	GH 10, 42, 35, LacA	35–36	2
3.	1045034	X_axonopodis-BgaX	<i>Xanthomonas axonopodis</i>	Prokaryota (bacteria)	598	GH42, 35, PLN03059, LacA	22–23	3
4.	32709094	X_campestris-GalC	<i>X. campestris</i>	Prokaryota (bacteria)	613	GH42, 35, PLN03059, LacA	23–24	2
5.	21114096	X_campestris-NixL	<i>X. campestris</i>	Prokaryota (bacteria)	613	GH42, 35, PLN03059, LacA	23–24	2
6.	2289790	B_circulans-BgaC	<i>Bacillus circulans</i>	Prokaryota (bacteria)	586	GH42, 35, PLN03059, BGal-dom4.5, LacA	No	2
7.	145688909	S_suis-BgaC	<i>Streptococcus suis</i> 05ZYH33	Prokaryota (bacteria)	590	GH42, 35, PLN03059, LacA	No	3
8.	14971525	S_pneumoniaeTIGR4-BgaC	<i>Streptococcus pneumoniae</i> TIGR4	Prokaryota (bacteria)	595	GH42, 35, PLN03059, LacA	No	2
9.	116077789	S_pneumoniaeD39-BgaC	<i>S. pneumoniae</i> D39	Prokaryota (bacteria)	595	GH42, 35, PLN03059, LacA	No	2
10.	15457592	S_pneumoniaeR6-BgaC	<i>S. pneumoniae</i> R6	Prokaryota (bacteria)	595	GH42, 35, PLN03059, LacA	No	2
11.	16611713	C_maltaromaticum-BgaC	<i>Carnobacterium maltaromaticum</i>	Prokaryota (bacteria)	586	GH42, 35, PLN03059, LacA	No	3
12.	257143787	P_thiaminolyticus-Bga	<i>Paenibacillus thiaminolyticus</i>	Prokaryota (bacteria)	583	GH42, 35, PLN03059, BGal-dom4.5, LacA	No	No
13.	669059	B_oleracea-BgalA	<i>Brassica oleracea</i>	Eukaryota (planta)	828	GH35, 42, PLN03059, Gal-Lectin, LacA	22–23	11
14.	68161828	M_indica-BGAL	<i>Mangifera indica</i>	Eukaryota (planta)	827	GH35, 42, PLN03059, Gal-lectin, LacA, DUF4185	22–23	7
15.	6686884	A_thaliana-BGAL6	<i>Arabidopsis thaliana</i>	Eukaryota (planta)	718	GH35, 42, PLN03059, BGal 4.5, LacA	28–29	3
16.	6686878	A_thaliana-BGAL3	<i>A. thaliana</i>	Eukaryota (planta)	856	GH35, 42, PLN03059, Gal-Lectin, LacA	No	2

(Continued)

TABLE 12.1 (Continued)

SI no.	GI number	Name used in the phylogenetic tree	Organisms	Taxonomy	No. of amino acids	Domains identified by CD-search	Signal peptide cleavage site	N-glycosylation sites
17.	20514290	O_sativa-BGAL1	<i>Oryza sativa</i>	Eukaryota (planta)	843	GH35, 42, PLN03059, BGal 4.5, Gal-lectin, LacA	No	2
18.	6686882	A_thaliana-BGAL5	<i>A. thaliana</i>	Eukaryota (planta)	732	GH35, 42, PLN03059, BGal 4.5, LacA	28–29	1
19.	3860321	C_arietinum-BGAL5	<i>Cicer arietinum</i>	Eukaryota (planta)	745	GH35, 42, PLN03059, LacA	26–27	1
20.	7682680	V_radiata-BGAL1	<i>Vigna radiata</i>	Eukaryota (planta)	739	GH35, 42, PLN03059, PRK13974, LacA	26–27	1
21.	56201401	R_sativus-BGAL1	<i>Raphanus sativus</i>	Eukaryota (planta)	851	GH35, 42, PLN03059, Gal_lectin, LacA	30–31	3
22.	14970841	F_X_ananassa-BGAL2	<i>Fragaria ananassa</i>	Eukaryota (planta)	840	GH35, 42, PLN03059, BGal 4.5, Gal_lectin, LacA	No	3
23.	7939623	S_lycopersicum-Tbg5	<i>Solanum lycopersicum</i>	Eukaryota (planta)	852	GH35, 42, PLN03059, Gal-lectin, LacA	No	5
24.	54291174	O_sativa-BGAL2	<i>O. sativa</i>	Eukaryota (planta)	715	GH35, 42, PLN03059, BGal 4.5, LacA	20–21	1
25.	20384648	C_sinensis-BGAL	<i>Citrus sinensis</i>	Eukaryota (planta)	737	GH35, 42, PLN03059, LacA	No	No
26.	452712	A_officinalis-BGAL	<i>Asparagus officinalis</i>	Eukaryota (planta)	832	GH35, 42, PLN03059, Gal-Lectin, LacA	25–26	No
27.	3641865	C_arietinum-BGAL4	<i>C. arietinum</i>	Eukaryota (planta)	723	GH35, 42, PLN03059, LacA	23–24	4
28.	3869280	C_papaya-BGAL	<i>Carica papaya</i>	Eukaryota (planta)	721	GH35, 42, PLN03059, LacA	21–22	1
29.	18148449	P_americana-BGAL1	<i>Persea americana</i>	Eukaryota (planta)	766	GH35, 42, PLN03059, LacA	35–36	1
30.	13936236	C_annuum-BGAL1	<i>Capsicum annuum</i>	Eukaryota (planta)	724	GH35, 42, PLN03059, LacA	23–24	3
31.	3299896	S_lycopersicum-Tbg4	<i>S. lycopersicum</i>	Eukaryota (planta)	724	GH35, 42, PLN03059, LacA	23–24	3
32.	4138137	S_lycopersicum-Tbg3	<i>S. lycopersicum</i>	Eukaryota (planta)	838	GH35, 42, PLN03059, Gal-lectin, LacA	25–26	1
33.	6649906	S_lycopersicum-Tbg1	<i>S. lycopersicum</i>	Eukaryota (planta)	835	GH35, 42, PLN03059, Gal-lectin, LacA	22–23	2
34.	14970839	F_X_ananassa-BGAL1	<i>F. ananassa</i>	Eukaryota (planta)	843	GH35, 42, PLN03059, Gal-lectin, LacA	28–29	2

35.	33521214	S_aurantiaca-BGAL	<i>Sandersonia aurantiaca</i>	Eukaryota (planta)	826	GH35, 42, PLN03059, Gal-lectin, LacA	24–25	2
36.	9294020	A.thaliana-BGAL1	<i>A. thaliana</i>	Eukaryota (planta)	847	GH35, 42, PLN03059, Gal-lectin, LacA	32–33	No
37.	14970843	F_X_ananassa-BGAL3	<i>F. ananassa</i>	Eukaryota (planta)	722	GH35, 42, PLN03059, LacA	25–26	1
38.	7682677	V_radiata-BGAL2	<i>V. radiata</i>	Eukaryota (planta)	721	GH35, 42, PLN03059, LacA	23–24	2
39.	10059008	D_caryophyllus-BGAL	<i>Dianthus caryophyllus</i>	Eukaryota (planta)	731	GH35, 42, PLN03059, LacA	No	2
40.	3860420	L_angustifolius-BGAL	<i>Lupinus angustifolius</i>	Eukaryota (planta)	730	GH35, 42, PLN03059, LacA	33–34	1
41.	507278	M_domestica-BGAL	<i>Malus domestica</i>	Eukaryota (planta)	731	GH35, 42, PLN03059, LacA	24–25	1
42.	12583687	P_pyrifolia-BGAL1	<i>Pyrus pyrifolia</i>	Eukaryota (planta)	731	GH35, 42, PLN03059, LacA	24–25	1
43.	8809655	A_thaliana-BGAL4	<i>A. thaliana</i>	Eukaryota (planta)	724	GH35, 42, PLN03059, LacA	27–28	1
44.	6686876	A_thaliana-BGAL2	<i>A. thaliana</i>	Eukaryota (planta)	727	GH35, 42, PLN03059, LacA	27–28	1
45.	334305536	L_usitatissimum-BGAL	<i>Linum usitatissimum</i>	Eukaryota (planta)	731	GH35, 42, PLN03059, LacA	29–30	1
46.	7939621	S_lycopersicum-Tbg7	<i>S. lycopersicum</i>	Eukaryota (planta)	870	GH35, 42, PLN03059, Gal-lectin, LacA	35–36	5
47.	6686892	A_thaliana-BGAL10	<i>A. thaliana</i>	Eukaryota (planta)	741	GH35, 42, PLN03059, LacA	29–30	4
48.	219927064	T_majus BGAL	<i>Tropaeolum majus</i>	Eukaryota (planta)	857	GH35, 42, PLN03059, Gal-lectin, LacA	No	7
49.	3641863	C_arietinum-BGAL3	<i>C. arietinum</i>	Eukaryota (planta)	730	GH35, 42, PLN03059, LacA	No	1
50.	18958133	A_candidus-BGAL	<i>Aspergillus candidus</i>	(Fungi)	1005	GH35, PLN03059, BGal-dom 2, 3, 4.5, 4.5, LacA	18–19	6
51.	582890099	A_oryzae-BGAL1	<i>Aspergillus oryzae-112</i>	(Fungi)	1005	GH35, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	18–19	6
52.	83770489	A_oryzae-BGAL2	<i>A. oryzae-RIB40</i>	(Fungi)	1005	GH35, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	18–19	6
53.	34370136	<i>Trichoderma reesei</i>	<i>T. reesei</i>	(Fungi)	1023	GH35, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	20–21	6
54.	321150462	P_aerugineus-bglA	<i>Paecilomyces aerugineus</i>	(Fungi)	1011	GH35, PLN03059, GH_2 N, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	18–19	6
55.	189092779	P_expansum-BGAL	<i>Penicillium expansum</i>	(Fungi)	1011	GH35, PLN03059, GH_2N, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	18–19	6

(Continued)

TABLE 12.1 (Continued)

SI no.	GI number	Name used in the phylogenetic tree	Organisms	Taxonomy	No. of amino acids	Domains identified by CD-search	Signal peptide cleavage site	N-glycosylation sites
56.	56266627	P_canescens-BGAL	<i>Penicillium canescens</i>	(Fungi)	1011	GH35, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	19–20	7
57.	44844271	P_sp.-BGAL	<i>Penicillium</i> sp.	(Fungi)	1011	GH35, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	19–20	6
58.	238914608	B_sp.MEY-1-bglA	<i>Bispora</i> sp. <i>MEY-1</i>	(Fungi)	1002	GH35, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	21–22	10
59.	32448796	R_emersonii-BGAL	<i>Rasamsonia emersonii</i>	(Fungi)	1008	GH35, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	19–20	9
60.	166513	A_niger-BGALA	<i>Aspergillus niger</i>	(Fungi)	1006	GH35, PLN03059, GH_2N, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	18–19	11
61.	62913951	A_phoenicis-BGAL	<i>Aspergillus phoenicis</i>	(Fungi)	1007	GH35, PLN03059, GH_2N, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	18–19	11
62.	383212688	P_chrysogenum-BGAL	<i>Penicillium chrysogenum</i>	(Fungi)	1013	GH35, 42, PLN03059, BGal-dom 2, 2, 3, 4.5, 4.5, LacA	21–22	7
63.	14099962	Cl_familiaris-BGAL	<i>Canis lupus familiaris</i>	(Primates)	668	GH35, 42, PLN03059, BGal-dom4.5, LacA	24–25	6
64.	192187	M_musculus-BGAL	<i>Mus musculus</i>	(Primates)	647	GH35, 42, PLN03059, BGal-dom4_5, LacA	24–25	6
65.	179401	H_sapiens-BGAL1	<i>Homo sapiens</i>	(Primates)	677	GH35, 42, PLN03059, BGal-dom4.5, LacA	23–24	7
66.	2547317	F_catus-BGAL	<i>Felis catus</i>	(Primates)	669	GH35, 42, PLN03059, BGal-dom4.5, LacA	24–25	6
67.	34013388	T_kodakarensis	<i>Thermococcus kodakarensis</i>	Archea	786	GH42, 35, PLN03059, LacA, A4 galactosidase middle domain, GH42 trimerization domain		
	NO	1						

BGal_dom 2, 2, 3, 4_5, 4_5, Beta-galactosidase domain 2, 2, 3, 4_5, 4_5; *GH_2N*, glycosyl hydrolase 2N sugar binding domain; *GH35*, glycosyl hydrolase-35; *GH42*, glycosyl hydrolase 42; *LacA*, beta-galactosidase; *PLN03059*, provisional multi domain; *PRK13974*, thymidylate kinase.



FIGURE 12.1 Conserved domains in mango BGAL were searched by Conserved Domain Database search in NCBI-BLAST. BGAL, Beta-galactosidase.

The signal peptide is a short peptide found in newly synthesized protein at *N*-terminal, which determines whether the protein will be secreted or not. The online web server “SignalP 4.1” was used to predict the signal peptide in the 67 BGAL amino-acid sequences (Petersen, Brunak, Heijne, & Nielsen, 2011). Fifty-one out of 67 BGALs possessed signal sequences in their polypeptide chains. Plant and animal BGAL signal peptides contain the first 21–35 amino acids, whereas fungal BGALs have 18–22 amino acids. MiBGAL signal peptide was found to be the first 23 amino acids (Table 12.1) (Hossain et al., 2016). Thirteen out of 17 BGALs were found in *Arabidopsis* that have potential *N*-terminal signal peptides secreted to the endomembrane system. The rest of the BGALs are probably located in the cytoplasm or nucleus (Chandrasekar & Hoorn, 2016). With the few exceptions in some bacteria, most of the BGALs have signal peptides in their polypeptide chains, indicating that they possibly are secreted proteins (Hossain et al., 2016).

Glycosylation is one of the most abundant posttranslational modification events in eukaryotes. The online web server “NetNGlyc 1.0” (<http://www.cbs.dtu.dk/services/NetNGlyc/>) is usually used to determine the potential *N*-glycosylation sites in the polypeptide sequences. The BGALs belonging to animals and fungi have 6–11 *N*-glycosylation sites. Most of the pBGALs contain less *N*-glycosylation sites than fungi. However, seven potential *N*-glycosylation sites were found in MiBGAL protein sequence. They are located at positions N24, N152, N252, N349, N378, N492, and N498 (Hossain et al., 2016). Most of the bacteria contained two *N*-glycosylation sites but don’t have any signal peptide, indicating that bacterial BGALs are not true glycosylation site (Table 12.1). Some bacteria have multifunctional proteins that are glycosylated and secreted or surface-exposed and might have an important role in the interaction with their environment (Szymanski & Wren, 2005). On other hand, *Penicillium* sp. BGAL contains seven *N*-linked oligosaccharide chains and was reported to be the first X-ray solved crystal structure of a glycosylated β -galactosidase (Rojas et al., 2004). Human BGAL also contains seven *N*-glycosylation sites at positions N26, N247, N464, N498, N542, N545, and N555 (Ohto et al., 2012). Two *N*-glycosylation sites (N282 and N459) have been reported in *Solanum lycopersicum* β -galactosidase 4 (TBG4), and a peptide signal cleavage site is found in between the amino-acid position of 23 and 24 in polypeptide sequence, indicating a high probability for secretory nature of protein (Hossain et al., 2016).

Usually, MEME online software is used to identify the conserved motif in the protein sequences (Bailey et al., 2009). Five conserved motifs are present in the 67 BGALs of plant, animals, fungi, and bacteria (Fig. 12.2). The number of amino-acid present in the motif-1, -2, -3, -4, and -5 are 50, 41, 41, 21, and 21, respectively (Hossain et al., 2016). Most of the BGAL sequences possess at least three common motifs: motif-1 (cyan), -4 (pink), and -5 (yellow) (Fig. 12.3) (Hossain et al., 2016). All pBGALs contain 5 motifs present in the domain-I (TIM barrel) and some have more than one copy of the same motif. It has been reported that it could be due to segmental gene duplication in the pBGALs (Hossain et al., 2016). A special motif 3 (red) is found at the active site of all pBGALs belonging to GH35 family. The motif 2 (blue) is also reported at the active site of pBGALs and bacterial that belongs to GH42. On the other hand, fungi don’t possess motif 2 (blue) whereas animals and bacterial BGALs also don’t have motif 3 (red) due to the short sizes of protein sequences. However, two animal BGALs have two copies of motif-1. No conserved motif was found at the C-terminal end of all BGAL polypeptide sequences, probably due to a lower similarity score among the member sequences (Hossain et al., 2016). Another online software, “PROSITE” (<http://www.prosite.expasy.org>) identified the GH35 predictive active site in pBGAL polypeptide sequences, which possesses a consensus sequence G-G-P-[LIVM](2)-x(2)-Q-x-E-N-E-[FY]. It was postulated that the second E was the key residue in the active site of pBGALs. More than 50% pBGALs contain the SUEL-type lectin domain (Hossain et al., 2016).

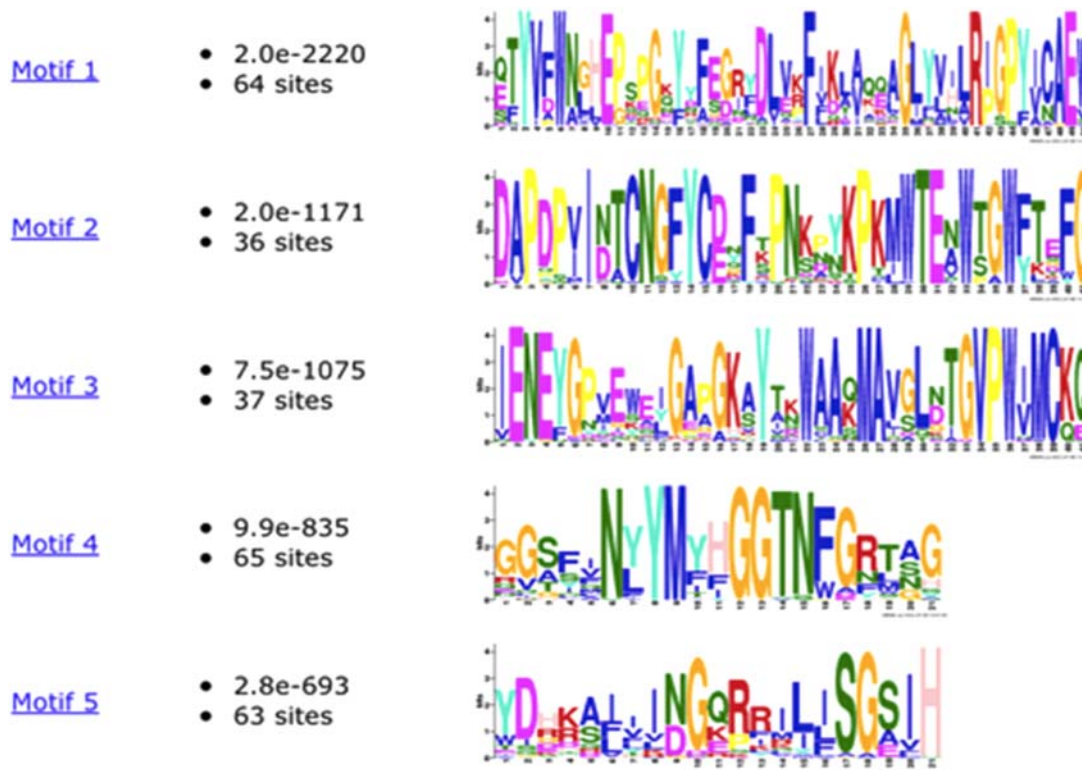


FIGURE 12.2 Conserved motifs present in the 67 BGALs protein sequences. Five motifs were identified using MEME (Motif Em for Motif Elicitation) software. The symbol heights represent the relative frequency of each residue. The number of sites and e-value for each motif are indicated. The widths of the motif-1, -2, -3, -4, and -5 are 50, 41, 41, 21, and 21 amino acids, respectively. *BGAL*, Beta-galactosidase.

12.3 Molecular evolution of beta-galactosidases and their classification

To determine the evolutionary relationship of various BGALs, a phylogenetic tree was reconstructed using 67 BGAL protein sequences retrieved from CAZy database. The MUSCLE program (Edgar, 2004) was used to alignment protein sequences. The phylogenetic relationships were built by PROTML program of PHYLIP version 3.6 using the maximum-likelihood method (Felsenstein, 2000). The BGAL genes evolved from an archea and organized into four different families such as bacteria, animals, fungi, and plants (Fig. 12.3) (Hossain et al., 2016). Further plants BGALs are subdivided into six subfamilies (D1–D6) where MiBGAL belongs to the D1 family. Bacterial BGAL proteins have the highest similarities to animals, and pBGALs evolve from fungi (Hossain et al., 2016). Approximately 65.57% of similarity index (identities) is found between MiBGAL and the *Brassica oleracea*-BGAL. The D5 subfamily members had the highest percentage (98.92%–65.77%) of sequence identities, whereas the D2 subfamily members had the lowest percentage (71.76%–47.16%). Although pBGALs have a wide range of protein sequence variation, all of them possess five conserved motifs, motif-1, -2, -3, -4, and -5 (Fig. 12.3). Few pBGALs have double motifs, which may be due to the segmental gene duplication events. It has been reported that all organisms except plants have a single copy of BGAL gene located in their chromosomes (Hossain et al., 2016). All plant species have multiple copies of BGAL genes, namely, 17 in *Arabidopsis* and tomato (Chandrasekar & Hoorn, 2016), 15 in rice (Tanthanuch et al., 2008), and 16 in brassica (Liu et al., 2013). The pBGAL multigenes reside either on the same or different chromosomal locations and they possibly evolved through segmental or gene duplications. The gene duplication might have critical roles in evolving new functions of the multifunctional enzymes (Hossain & Roslan, 2014).

Protein sequence analyses reveal that pBGALs can be divided into two subgroups based on their length of polypeptide chain; smaller BGALs (Less than 750 aa) possess GH35, 42, and β -galactosidase domains (Hossain et al., 2016). Larger BGALs (Greater than 750 aa) contain the conserved C-terminal Lectin-like SUE (sea urchin egg lectin) type domains. Lectin-like SUE domains usually contain 100 amino acids with 7 highly conserved cysteine residues. This C-terminal domain is very common to many pBGALs, which shows homology to animal lectin proteins (Ozeki, Yokota, Kato, Titani, & Matsui, 1995). Sea urchin eggs also contain SUE lectins, which consist of L-rhamnose- and D-galactose-

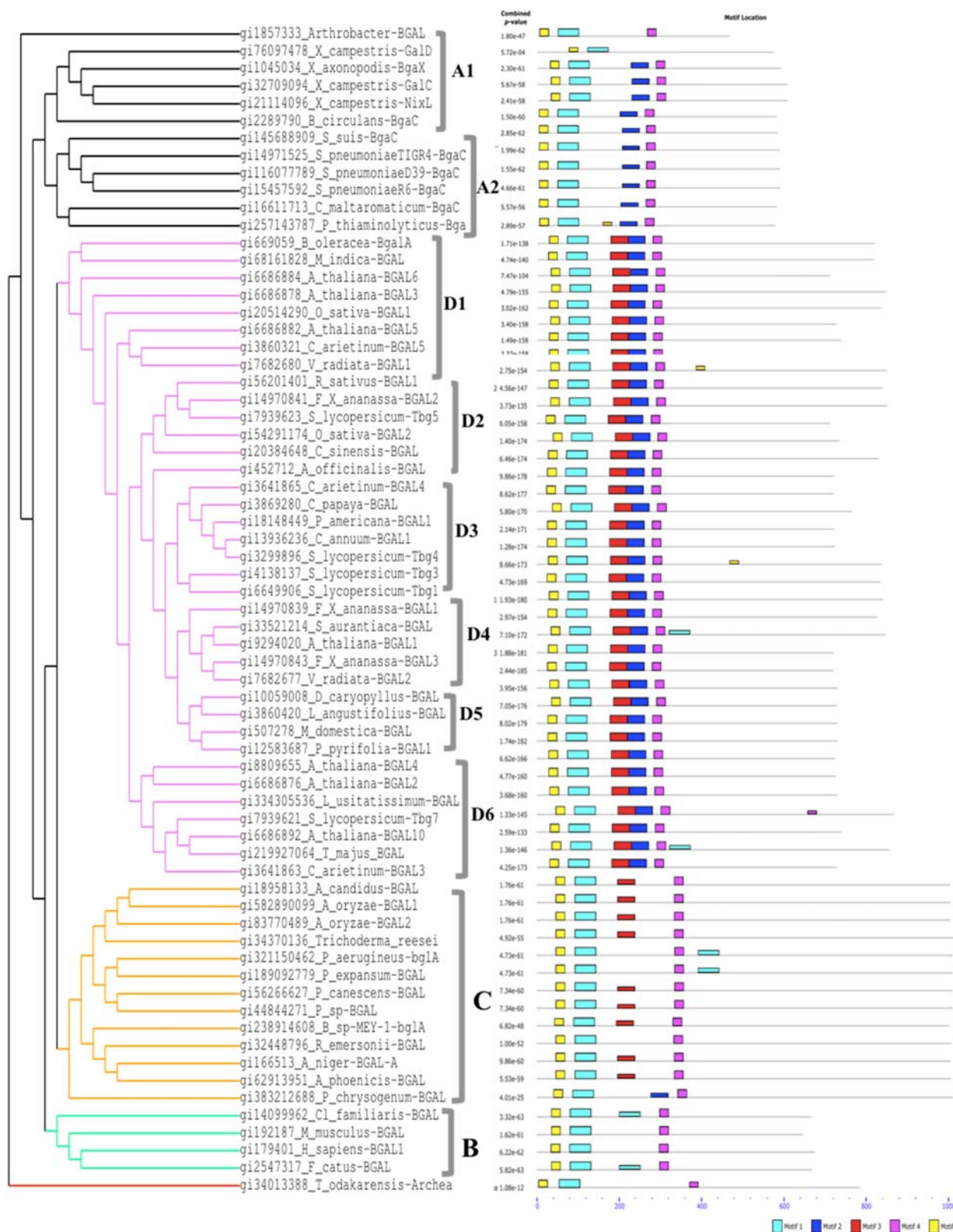


FIGURE 12.3 The phylogenetic tree (cladogram) based on beta-galactosidase (GH35) amino-acid sequences obtained by the maximum-likelihood method (left side). Archaean GH35 sequences were used as an out-group to reconstruct the phylogenetic tree. All analyses were performed with the WAG amino-acid substitution model and 1 invariable and 4 gamma-distributed site rate categories. Detailed information about the sequences is shown in Table 12.1. The conserved motifs are distributed in the BGALs sequences (right side).

specific homodimers (Ozeki et al., 1995). Although there are no experimental evidences on the specific function of this domain in plants (Tanthanuch et al., 2008), it has been suggested that the lectin-like domain could enhance the affinity of the enzymes for their substrates, thereby increasing catalytic efficiency (Ahn et al., 2007) and possibly also enzyme stability (Trainotti, Spinello, Piovan, Spolaore, & Casadoro, 2001).

12.4 Three-dimensional structural characteristics of plant beta-galactosidases

3D structure is very important for the determination of structure–function relationship of the proteins and/or enzymes. Still now only tomato pBGAL (TGB4) structure has been successfully solved by X-ray crystallography (Eda et al., 2015). An open reading frame of TGBG4 cDNA (24–724 aa) was cloned and expressed in *Pichia pastoris* using expression vector pPICZ_A (Invitrogen) after the α -factor signal sequence for the production of a secreted recombinant protein fused with a hexahistidine tag (Eda et al., 2015). BGAL isolated from different sources has a common catalytic TIM-barrel domain in their structure. The catalytic domain of TGB4 showed amino-acid sequence identities at 27%–34% with other enzymes and the other part of the TGB4 have 19%–25% sequence identities (Eda et al., 2015). The Ramachandran plot shows that over 95.9% of the residues in the crystal structures remain in structural favor region and 3.9% residues in the allowed region, whereas only 0.2% residues fall in disallowed region (Masahiro Eda et al., 2016). These results indicate that the structure is of high quality and accurate. TGB4 consists of four domains (Masahiro Eda et al., 2016); a central TIM-barrel domain is followed by three β -sandwich domains (Fig. 12.4A). The domain-I has 323 amino acids (24–346) with a distorted (β/α)8 TIM barrel fold that houses the active site. Like an ideal TIM barrel of 8 (β/α) repeats, the TGB4 TIM barrel does not have the fifth and sixth α -helices in the β/α barrel (Masahiro Eda et al., 2016). The domain-II contains 66 amino-acid residues (347–412) with an antiparallel β -sandwich structure that possesses 7 β -strands (Masahiro Eda et al., 2016). The domain-III contains an antiparallel β -sandwich structure of 9 β -strands joined with the loop structure. Amino-acid residues located at the position 413–438 in the polypeptide chain build up loop regions and 2 β -strands, and residues 586–724 constitute the rest of the C-terminal domain (Masahiro Eda et al., 2016). The domain-IV contains 147 amino acids (439–585) that form an antiparallel β -sandwich structure with 8 β -strands (Masahiro Eda et al., 2016).

To get more insights of substrate specificities, mechanism of action pBGALs and physiological function, a modeled structure of MiBGAL also has been developed by homology modeling using TGB4 as a template (pdb id: 3w5g) (Masahiro Eda et al., 2016). Homology modeling was carried out on online software, SWISS-MODEL web server (Waterhouse et al., 2018) followed by ModRefiner (Xu & Zhang, 2011). An important criterion of reliable homology modeling is the cut off value greater than 30% sequence identity between the template and target. More importantly,

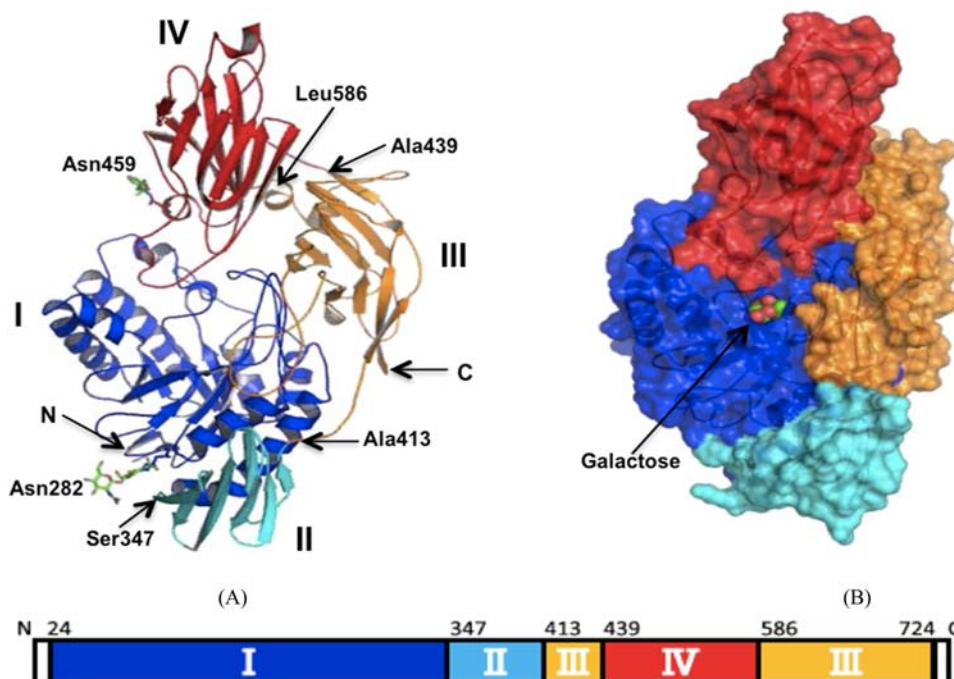


FIGURE 12.4 Three-dimensional X-ray solved crystal structure of (A) TGB4 and (B) its complex with galactose. Four domains I–IV are colored blue, cyan, orange, and red, respectively. Glycosylated amino-acid residues (Asn282 and Asn459) and N-acetyl-D-glucosamine residues are depicted as stick models in (A). The β -D-galactose molecule in the active site is shown as a space-filling view in (B).

more than 50% sequence identity between the template and target usually give an accurate model. The amino-acid sequence identity of MiBGAL with TBG4 is 52.80% (Hossain et al., 2016). The modeled structure is analyzed using the protein structure and model assessment tools at the SWISS-MODEL server, which utilizes various local and global quality estimation parameters. Finally, the model is assessed and verified using the PROCHECK (Laskowski, MacArthur, & Thornton, 2001), WHAT_CHECK (Hoof, Vriend, Sander, & Abola, 1996), VERIFY_3D (Luthy, Bowie, & Eisenberg, 1992) methods, and ModEval Model evaluation server (Eramian, Eswar, Shen, & Sali, 2008). Over 99.5% of the residues fell in the common region in the Ramachandran plot and only 0.50% remained in the unfavorable region, indicating that the refined structure is good quality (Hossain and Roslan, 2014). The overall G-factor, the main chain, side chains, and bond angles parameters were found within the normal limits.

Since the template structure, TBG4 (pdb id: 3w5g) was cloned and expressed *P. pastoris* without signal sequence (Eda et al., 2015). Like the template structure, the modeled structure of MiBGAL starts at position 22 and included 716 amino acids (position 22–737) in its structure (Hossain et al., 2016). The MiBGAL modeled 3D structure is presented in Fig. 12.5A. Similar to TBG4, the model contains four domains (Hossain et al., 2016); first domain (GH35) comprises a triose-phosphate isomerase (TIM) barrel [also called (β/α)₈] in the central part houses the catalytic residues, second and third domains form small beta-sandwich and fourth domain is jelly-roll like (Fig. 12.5A and B) as found in template structure, TBG4 (Hossain et al., 2016). The second, third, and fourth domains consist of six, seven, and eight antiparallel β -sandwich structures, respectively (Hossain et al., 2016). The secondary-structural elements of the modeled structure MiBGAL consist of 10 α -helices and 38 β -strands, with two additional disulfide bonds located at the position C230–C235 and C372–405 (Hossain et al., 2016). PROMOTIF predicted the α - and β -contents of the modeled structure that are 15.90% and 29.10%, respectively. Superimposition of modeled MiBGAL with template TBG4 exhibits the magic-fit overlapping conformation (Fig. 12.4E) (Hossain et al., 2016). It has been reported that five distinct domains are found in *Penicillium* sp. BGAL (PspBGAL) that belongs to GH35 (Rojas et al., 2004). The PspBGAL has two disulfide bonds at the position of C205–C206 and C267–C316.

12.5 Structural comparison between MiBGAL and TBG4

Both the MiBGAL modeled and TBG4 crystal-structure possess the four domains including TIM barrel in their centers (Fig. 12.5A–D) (Hossain et al., 2016). The active site clefts of both structures are also very similar in conformation (Fig. 12.5B and D) (Hossain et al., 2016). The COACH program identified the catalytic residues responsible for the ligand binding of the modeled structure of MiBGAL. Thirteen interacting residues located at the catalytic site of modeled MiBGAL structure are Tyr74, Val117, Cys118, Ala119, Glu120, Asn181, Glu182, Glu251, Trp253, Trp256, Phe257, Tyr290, and Tyr313, whereas that residues of TBG4 are Tyr74, Val117, Cys118, Ala119, Glu120, Asn180, Glu181, Glu250, Trp252, Trp255, Tyr256, Tyr289, and Tyr312 (Hossain et al., 2016). The interacting residues are very much similar in both structures. Superimposition of MiBGAL (model) with TBG4 structure presents a magic fit between them (Fig. 12.5E) and the only difference is that the MiBGAL contain Phe257 instead of Tyr256, which is located in the catalytic site of TBG4 crystal structure (Fig. 12.5F) (Eda et al., 2015). The MiBGAL also possesses two disulfide bonds at the position of C230–C235 and C372–C405, whereas TBG4 structure forms four disulfide bonds at the position of C229–C234, C370–C405, C684–C682, and C67–C678 (Masahiro Eda et al., 2016). Protein–protein interaction studies revealed that MiBGAL exists in dimeric form as found in TBG4. The galactose–TBG4 protein interaction (pdb id: 3w5g) is represented in Fig. 12.5A, where the conserved amino acids residues such as Tyr74, Cys118, Ala119, Glu120, Asn180, Glu181, Asn230, Glu250, Trp252, Tyr289, and Tyr312 of TBG4 interacted with galactose molecule to form complex (Hossain et al., 2016). Superposition of the PspBGAL–galactose complex with other BGAL complexes belonging to GH35 identified the E200 and E299 residues as the proton donor and the nucleophile, respectively (Rojas et al., 2004). The Glu182 and Glu251 were identified as catalytic residues in a deep well in the TIM-barrel domain of MiBGAL modeled structure by triple-superimposition (Hossain et al., 2016). The residue Glu182 and Glu251 could act as the proton donor the catalytic nucleophile base of MiBGAL (Fig. 12.6B). On the other hand, the Glu181 and Glu250 were identified as proton donor and catalytic nucleophile in TBG4, respectively (Fig. 12.6A and B) (Eda et al., 2015).

12.6 Substrate specificity of plant beta-galactosidases

The pBGAL can hydrolyze various plant-based (1,4)-linked polysaccharides and exhibits a strong affinity to attack β -(1,4)-galactan molecules. Three aromatic amino-acid residues postulated to be important for substrate specificity are conserved in GH35 BGALs isolated from bacteria, fungi, and animals (Cheng et al., 2012). The crystal structural

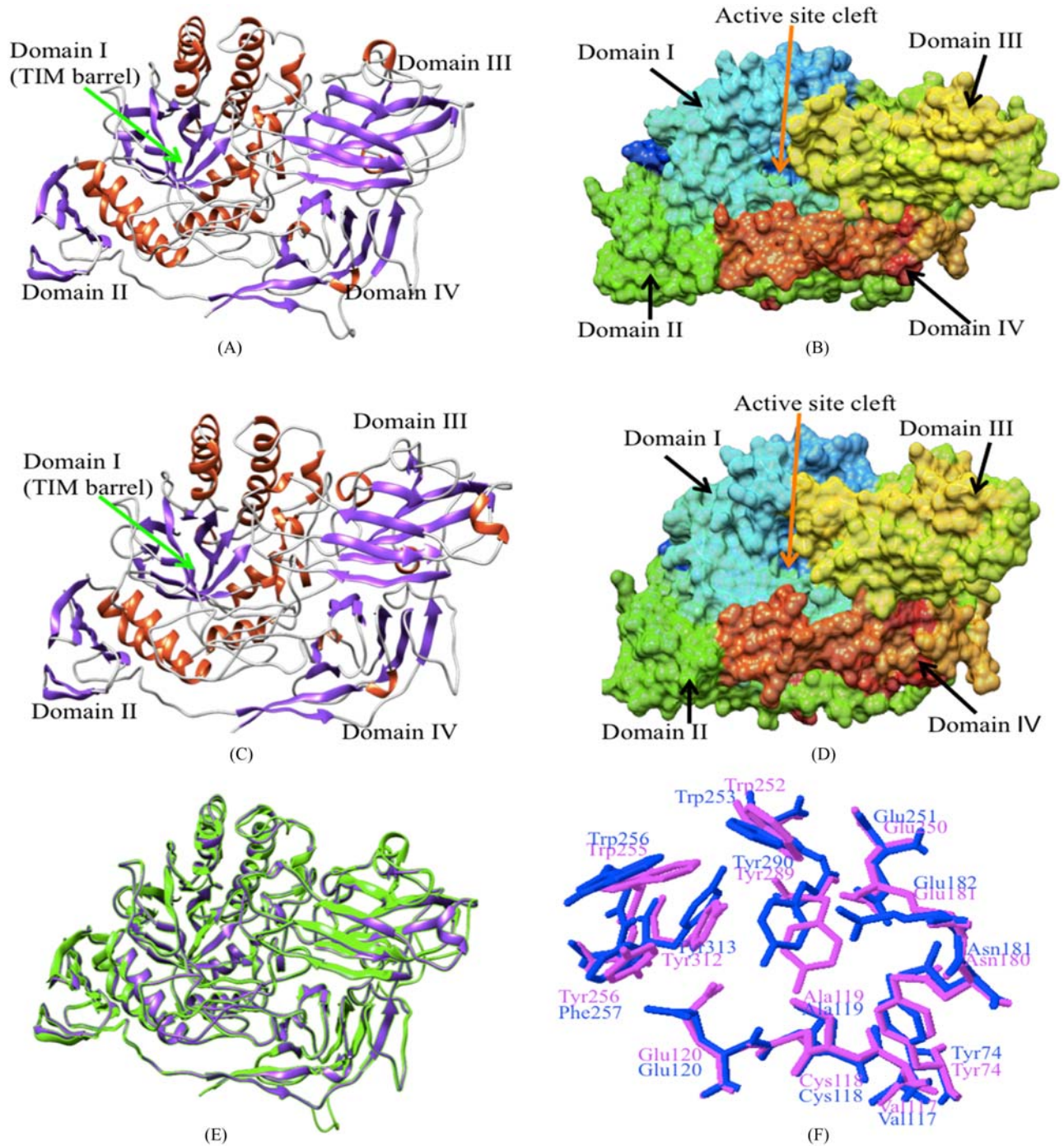
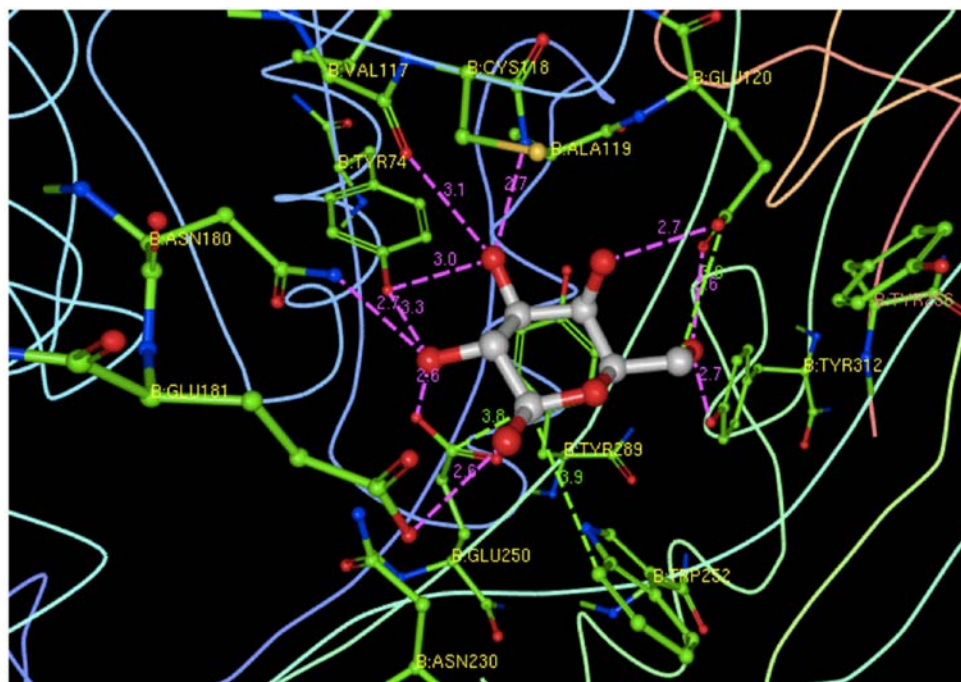
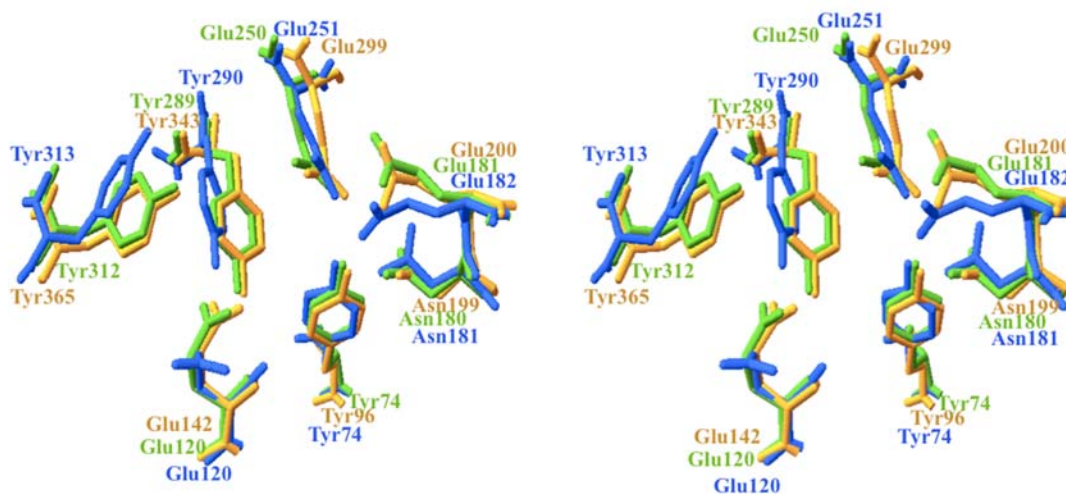


FIGURE 12.5 The molecular 3D structural features of plant beta-galactosidases (BGALs). (A) The predicted 3D-modeled structure of mango BGAL (MiBGAL) is shown as ribbon diagram. The structure contains fourfold domains (I, II, III & IV) including α -helices (red), β -pleated sheets (purple), and coils (gray) as found in template structure (pdb id: 3w5g). The catalytic domain-I is a TIM barrel with the active site located at the *N*-terminus of the protein. (B) Surface (20% transparent) of modeled structure of MiBGAL. Four domains are depicted as cyan (domain-I), green (domain-II), yellow (domain-III), red (domain-IV), and active site clefts are indicated by red arrows. Parts (C) and (D) are the ribbon forms and surface (20% transparent) filled view of X-ray crystal structures of tomato BGAL (TBG4), respectively. (E) Superimposition magic-fit image of the modeled structures of MiBGAL (Green) with template structure TBG4 (Id: 3w5g, purple). (F) Superimposition of magic-fit image of active site residues of modeled MiBGAL (blue) with template, 3w5g (magenta). Chimera and SPDBV 4.01 OSX were used to prepare the images.



(A)



(B)

FIGURE 12.6 (A) Representations of the galactose-protein interactions in the catalytic sites of the X-ray crystal structure of tomato beta-galactosidase-galactose complex (TBG4; pdb id: 3w5g). Bonds and bond lengths are indicated as purple (H-bonds), green (hydrophobic bonds). Bond lengths are expressed as Angstrom (\AA). (B) Stereo view of the superposition of the catalytic sites in MiBGAL (in blue), TBG4 (in green), and PspbGAL (in orange). The pictures were prepared by RCSB-PDB Ligand explorer 4.2.0 and SWISS-PDB viewer.

complex of TBG4 with β -D-galactose ligand provided structural insight into its substrate specificity (Eda et al., 2015). In TBG4 structure, two amino acids out of three were conserved and one aromatic residue was replaced by valine residue. To confirm the role of valine residue, kinetic studies were carried out of TBG4 and its mutant V548W using the synthetic (4-nitrophenyl β -D-galactopyranoside) and natural substrates [β (1,4), β (1,3), and β (1,6)-galactobiose, chelator-soluble pectin, alkali-soluble pectin] (Ohto et al., 2012). It is interesting that V548W mutant showed fivefold more activity (K_{cat} value) compared with wild-type TBG4, but k_{m} values remained the same levels for both (Ohto et al., 2012).

TABLE 12.2 Molecular docking for substrates of plant β -galactosidases (Hossain et al., 2016).

Sl no.	Name of ligands	Mango BGAL modeled structure		TBG4 X-ray crystal structure	
		Rosetta Interface Energy (delta_X score)	Auto Dock Binding free energy (ΔG), kcal/mol	Rosetta Interface Energy (delta_X score)	Auto Dock Binding free energy (ΔG), kcal/mol
	Substrates				
1.	p-Nitrophenyl- β -D-galactopyranoside (pNP-GAL)	-15.20	-5.18	-11.30	-4.50
2.	o-Nitrophenyl- β -D-galactopyranoside (oNP-GAL)	-12.29	-5.09	-8.79	-3.81
3.	(+)Lactose	-8.39	-3.94	-8.48	-3.26
4.	Galactobiose	-9.29	-4.38	-7.31	-3.45
5.	Galactotriose	Not possible	-2.88	Not possible	+0.08
6.	Arabinogalactan	Not possible	-0.80	Not possible	-1.31
7.	Galactan	Not possible	+3.71	Not possible	+5.79

The activity of the V548W mutant as compared with wild-type TBG4 increased sixfold against β -(1,6)-galactobiose and \sim 0.6-fold against β -(1,4)-galactobiose, while no change of activity against β -(1,3)-galactobiose (Ohto et al., 2012). The V548W mutant hydrolyzed the chelator-soluble pectin and alkali-soluble pectin and released the galactose molecule approximately 0.6–0.8-fold compared with wild-type TBG4, indicating that V548 have a critical role in substrate specificity and efficiently degrade the pectic β (1,4)-galactan (Ohto et al., 2012). Another report showed that TBG5, which had tyrosine residue instead of valine at same of TBG4, preferred to hydrolyze β (1,6) and β (1,3)-linked galactooligosaccharides but did not show any activity against substrate, β (1,4)-galactan (Ishimaru, Smith, Mort, & Gross, 2009). Thus the residue present in the TBG4 corresponding to the position V548 of plant galactosidases seems to determine the substrate specificity against β (1,4)- and β (1,6)-linked polysaccharides (Ohto et al., 2012).

Molecular docking is the study of an interaction between the protein and ligands to determine the stability of the interacting amino-acid residues between the substrate and the active site (Meng, Zhang, Mezei, & Cui, 2011; Sethi, Joshi, Sasikala, & Alvala, 2019). We did molecular docking using 12 well-known synthetic and natural substrates as well as inhibitors (Tables 12.2 and 12.3). Molecular docking was carried out using the online server such as RosettaLigands (<http://rosettaserver.graylab.jhu.edu/>) and DockingServer (<http://www.dockingserver.com/>). In Rosetta docking the lowest Interface Energy (delta_X) scores were found in p-nitrophenyl- β -D-galactopyranoside (pNP-GAL) among the tested ligands for this study. The delta_X scores were -15.20 and -11.30 for the structure of MiBGAL and TBG4, respectively (Tables 12.2 and 12.3). Consistently, Auto Dock results showed that the binding free energies (ΔG) of pNP-GAL were -5.18 and -4.50 kcal/mol for MiBGAL and TBG4, respectively (Table 12.2). These results indicated that pNP-GAL could be potential synthetic substrate for both modeled structure of MiBGAL and the crystal structure of TBG4. It is also consistent with experimental results where TBG4 has strong activities toward the pNP-GAL (Smith and Gross, 2000). Therefore it could be concluded that pNP-GAL is potential synthetic substrate for both MiBGAL and TBG4; and MiBGAL might be able to hydrolyze β -(1 \rightarrow 4) linkage of the substrates like TBG4 (Ohto et al., 2012).

12.7 Mechanism of action of plant beta-galactosidases

BGALs isolated from different sources belonging to GH35 usually act to retain the same stereochemical configuration of product as initial substrate, which is called “retaining mechanism” (Rojas et al., 2004). Double-displacement reaction occurs here where two successive nucleophilic attack on the anomeric carbon that guides to overall retention of the anomeric configuration (Rojas et al., 2004). Two carboxylic acids are required for this reaction; First carboxylic acid

TABLE 12.3 Molecular docking for inhibitors of plant β -galactosidases (Hossain et al., 2016).

Sl no.	Name of ligands	Mango BGAL modeled structure		TBG4 X-ray crystal structure	
		Rosetta interface energy (delta_X score)	Auto dock binding free energy (ΔG), kcal/mol	Rosetta interface energy (delta_X score)	Auto dock binding free energy (ΔG), kcal/mol
	Inhibitors				
1.	1-Deoxy-manojirimycin (DM)	-9.68	-6.46	-9.47	-6.06
2.	1-Deoxy-galactonojirimycin (DG)	-7.91	-4.95	-9.28	-4.88
3.	Galactose	-9.63	-4.62	-9.19	-4.12
4.	1-methyl- β -D-galactoside	-7.15	-4.40	-8.76	-3.76
5.	1-methylcyclopropene	-4.86	-3.02	-7.97	-2.51

functions as a catalytic nucleophile, and second one works as an acid/base catalyst (Rojas et al., 2004; Ohto et al., 2012). In TBG4, Glu181 and Glu250 resided in the TIM-barrel domain were identified as a candidate for the acid/base catalyst and catalytic nucleophile, respectively (Ohto et al., 2012). A complex of galactose–TBG4 protein is shown in Fig. 12.6A, where a galactose is bonded with each monomer in the chair conformation, and its OH group in the position one exists in the β -anomeric form (Ohto et al., 2012). Nine hydroxyl groups of galactose form direct hydrogen bonds with TBG4 protein. In addition, aromatic and hydrophobic amino-acid residues resided at the catalytic site play an important role in the recognition of ligand through extensive van der Waals interactions (data not shown) (Hossain et al., 2016). The terminal galactose molecule is identified by TBG4 based on these interactions.

In the MiGBAL modeled structure, catalytic site conformation is very indistinguishable from TBG4 (Hossain et al., 2016). Two catalytically important residues, Glu182 and Glu251, are resided in the TIM-barrel domain and overlapped with identical residues upon triple-superimposition of MiBGAL, TBG4, and PspBGAL (Fig. 12.6B) (Hossain et al., 2016). Protein–ligand interaction (MiBFAL–Galactose) studies showed that the residue Glu182 and Glu251 could function as the proton donor and the catalytic nucleophile base of MiBGAL (data not shown) (Hossain et al., 2016). According to the conformation conservation of the anomeric carbon position through the reaction mechanism, glycoside hydrolases belonging to GH35 can be categorized into two types: retaining or inverting enzymes (Ohto et al., 2012). The basis of this classification is the distance between the oxygen of the two successive catalytic carboxylates; distance ranges for retaining enzyme and inverting enzyme are 4.5–6.5 and 9.0–9.5 Å, respectively (Ohto et al., 2012). The retaining enzyme reacts with its substrate to form a covalent intermediate, while the inverting enzyme hydrolyzes the substrate by activating the water molecules (Eda et al., 2015). The average value of the distance for the two carboxylates was 5.99 Å in the modeled structure of MiBGAL, while that of TBG4 was 5.41 Å, indicating that pBGALs act on their substrates in a retaining manner (Hossain et al., 2016).

12.8 Physiological function of plant beta-galactosidase

The pBGALs can hydrolyze β -(1,4)-galactans to play various physiological functions including cell-wall extension and breakdown of signaling molecules during fruit (Ross et al., 1993). The BGAL activity in tomato fruit significantly increased during ripening that suggested its roles in the breakdown of β (1,4)-galactan side chains of pectin as part of the ripening process (Carey et al., 1995). Another group of scientists reported that tomato β -galactosidase-4 (TBG4) hydrolyzes a wide varieties of plant-derived (1,4)- or 4-linked polysaccharides and exhibits a strong affinity to attack β -(1,4)-galactan, thereby expanding the cell-wall pericap (Eda et al., 2015). Recently, Yang's group reported that BGAL activity and expression levels of BGAL genes (Md β -Gal1, Md β -Gal2, and Md β -Gal5) significantly increased in

“Fuji” and “Qinguan” apples during all stages of fruit developmental and were much higher in the mature fruits, indicating that pectin was degraded by BGALs (Yang et al., 2018). Another report on pectin changes and pectin-modifying enzymes in Jonagold apples during postharvest softening showed that the BGAL was the key player for softening during ripening (Gwanpua et al., 2014).

Recently, the pBGALs have gained much interest for mostly their involvement in fruit developmental stages and pectin degradation during fruit ripening in various plants including tomato (Carey et al., 1995; Moctezuma et al., 2003; Pressey, 1983), muskmelon (Ranwala et al., 1992), kiwifruit (Ross et al., 1993), mango (Ali et al., 1995), peach (Lee et al., 2003), papaya (Lazan et al., 2004), and apple (Yang et al., 2018). Subsequently, it has been reported that downregulation of a ripening-related BGAL mRNA decreased the enzyme activity and freed galactose content and significantly retained the fruit firmness (Smith et al., 2002). Our previous study showed that mango ripening-related enzymes such as BGAL, α -mannosidase, and beta-hexosaminidase changed significantly during the postharvest storage at different temperatures (Hossain et al., 2014). The BGAL is thought to accelerate fruit softening by increasing the porosity of the cell wall and enhancing the access of other cell wall-degrading enzymes (Brummell, 2006; Ng et al., 2013, 2015). A β -galactosidase has been reported from chickpea (*Cicer arietinum*) seeds, indicating its involvement in plant seedling development (Kishore and Kayastha, 2012). Spinach leaf β -galactosidases also showed the synergistic action with α -L-arabinofuranosidases on the hydrolysis of arabinogalactan protein (Hirano, Tsumuraya, & Hashimoto, 1994). Recently, genome-wide identification and expression analysis revealed that sweet potato contains 17 BGAL genes that might be involved in plant development and stress responses through regulating the metabolism of cell-wall polysaccharides (Li et al., 2020). Although several reports have been published on BGALs found in various parts of plants such as fruits (Lazan et al., 2004; Lee et al., 2003), seeds (Kishore and Kayastha, 2012), and leaves (Hirano et al., 1994; Li et al., 2020), their physiological functions in the plant kingdom still remain obscure.

12.9 Conclusion

Evolutionary analyses revealed that all BGALs are evolved from the ancestor bacterial BGALs. All BGALs including plants have the most common TIM-barrel domains that house their catalytic residues. However, dissimilarities at the C-terminal region of the BGALs belonging to GH35 members are the cause of diversified or new functional of these enzymes different organisms. The pBGALs may function through a retaining mechanism as found in animal BGAL. Docking results showed clear pictures of the ligand-protein interactions and substrate specificities of pBGALs. Although X-ray crystal structure analyses of BGALs belonging to GH35 increase our understanding of the structure-function relationship, their exact roles of pBGLs in plant physiology remain elusive. To get a better understanding of the molecular functions of this enzyme in plant biology, it is advisable to characterize the properties, structures, and evolution of related BGALs from different species of plants.

Conflict of interest

The authors declare no conflict of interest.

References

- Ahn, Y. O., Zheng, M., Winkel, B., Bevan, D. R., Esen, A., Shin-Han, S., ... Shih, M. (2007). Functional genomic analysis of *Arabidopsis thaliana* glycoside hydrolase family 35. *Phytochemistry*, 68, 1510–1520.
- Ali, Z. M., Armugam, S., & Lazan, H. (1995). β -Galactosidase and its significance in ripening mango fruit. *Phytochemistry*, 38, 1109–1114.
- Bailey, T. L., Bodén, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... Noble, W. S. (2009). MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Research*, 37, W202–W208.
- Brummell, D. A. (2006). Cell wall disassembly in ripening fruit. *Functional Plant Biology*, 33, 103–119.
- Buckeridge, M. S., & Reid, J. S. (1994). Purification and properties of a novel β -galactosidase or *exo*-(1–4)- β -D-galactanase from the cotyledons of germinated *Lupinus angustifolius* L. seeds. *Planta*, 192, 502–511.
- Carey, A. T., Holt, K., Picard, S., Wilde, R., Tucker, G. A., Bird, C. R., ... Seymour, G. B. (1995). Tomato *exo*-(1–4)- β -D-galactanase. Isolation, changes during ripening in normal and mutant tomato fruit, and characterization of a related clone. *Plant Physiology*, 1008, 1099–1107.
- Chandrasekar, B., & Hoorn, R. A. L. (2016). Beta galactosidases in Arabidopsis and tomato—a mini review. *Biochemical Society Transactions*, 44, 150–157.
- Cheng, W., Wang, L., Jiang, Y. L., Bai, X. H., Chu, J., Li, Q., ... Chen, Y. (2012). Structural insights into the substrate specificity of *Streptococcus pneumoniae* β (1,3)-galactosidase BgaC. *Journal of Biological Chemistry*, 287, 22910–22918.

- de Alcantara, P. H., Martim, L., Silva, C. O., Dietrich, S. M., & Buckeridge, M. S. (2006). Purification of a β -galactosidase from cotyledons of *Hymenaea courbaril* L. (Leguminosae). Enzyme properties and biological function. *Plant Physiology and Biochemistry*, 44(11–22), 619–627.
- Dwevedi, A., & Kayastha, A. M. (2010). Plant β -galactosidases: Physiological significance and recent advances in technological applications. *Journal of Plant Biochemistry & Biotechnology*, 19(1), 09–20.
- Eda, M., Ishimaru, M., & Tada, T. (2015). Expression, purification, crystallization and preliminary X-ray crystallographic analysis of tomato β -galactosidase 4. *Acta Crystallography*, F71, 153–156.
- Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5, 113.
- Eramian, D., Eswar, N., Shen, M. Y., & Sali, A. (2008). How well can the accuracy of comparative protein structure models be predicted? *Protein Science*, 17, 1881–1893.
- Felsenstein, J. (2000). *PHYLIP: Phylogeny inference package, version 3.6 (alpha)*. Seattle, WA: University of Washington.
- Figueiredo, S. A., Lashermes, P., & Araga, F. J. L. (2011). Molecular characterization and functional analysis of the β -galactosidase gene during *Coffea arabica* (L.) fruit development. *Journal of Experimental Botany*, 62(8), 2691–2703.
- Gwanpua, S. G., Buggenhout, S. V., Verlinden, B. E., Christiaens, S., Shpigelman, A., Vicent, V., ... Geeraerd, A. (2014). Pectin modifications and the role of pectin-degrading enzymes during postharvest softening of Jonagold apples. *Food Chemistry*, 158, 283–291.
- Hirano, Y., Tsumuraya, Y., & Hashimoto, Y. (1994). Characterization of spinach leaf α -L arabinofuranosidases and β -galactosidases and their synergistic action on an endogenous arabinogalactan protein. *Physiologia Plantarum*, 92(2), 286–296.
- Hooft, R. W. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. *Nature*, 381, 272.
- Hossain, M. A., Rana, M. M., Kimura, Y., & Roslan, H. A. (2014). Changes in biochemical characteristics and activities of ripening associated enzymes in mango fruit during the storage at different temperatures. *BioMed Research International*, Article ID 232969, 11 pages.
- Hossain, M. A., & Roslan, H. A. (2014). Molecular phylogeny and predicted 3D structure of plant beta-D- β -acetylhexosaminidase. *The Scientific World Journal*, Article ID 186029, 14 pages.
- Hossain, M. A., Roslan, H. A., Karim, M. R., & Kimura, Y. (2016). Molecular phylogeny, 3D-structural insights, docking and mechanisms of action of plant beta-galactosidases. *International Journal of Bioinformatics Research and Applications*, 12(2), 149–179.
- Husain, Q. (2010). Beta-galactosidases and their potential applications: A review. *Critical Review of Biotechnology*, 30, 41–62.
- Ishimaru, M., Smith, D. L., Mort, A. J., & Gross, K. C. (2009). Enzymatic activity and substrate specificity of recombinant tomato β -galactosidases 4 and 5. *Planta*, 229, 447–456.
- Jacobson, R. H., Zhang, X. J., Dubose, R. F., & Matthews, B. W. (1994). Three-dimensional structure of β -galactosidase from *E. coli*. *Nature*, 369(6483), 761–766.
- Kishore, D., & Kayastha, A. M. (2012). A β -galactosidase from chickpea (*Cicer arietinum*) seeds: Its purification, biochemical properties and industrial applications. *Food Chemistry*, 13, 1113–1122.
- Kotake, T., Dina, S., Konishi, T., Kaneko, S., Igarashi, K., Samejima, M., et al. (2005). Molecular cloning of a β -galactosidase from radish that specifically hydrolyzes β -(1→3)- and β -(1→6)-galactosyl residues of arabinogalactan protein. *Journal of Plant Physiology*, 138, 1563–1576.
- Laskowski, R. A., MacArthur, M. W., & Thornton, J. M. (2001). PROCHECK: Validation of protein structure coordinates. In M. G. Rossmann, & E. Arnold (Eds.), *International tables of crystallography, Volume F. Crystallography of biological macromolecules* (pp. 722–725). The Netherlands: Dordrecht, Kluwer Academic Publishers.
- Lazan, H., Ng, S.-Y., Goh, L.-Y., & Ali, Z. M. (2004). Papaya β -galactosidase/galactanase isoforms in differential cell wall hydrolysis and fruit softening during ripening. *Plant Physiology and Biochemistry*, 42–847–853.
- Lee, D. H., Kang, S.-G., Suh, S.-G., & Byun, J. K. (2003). Purification and characterization of a β -galactosidase from peach (*Prunus persica*). *Molecular Cells*, 15(1), 68–74.
- Li, Z., Hou, F., Du, T., Xu, T., Li, A., Dong, S., ... Zhang, L. (2020). Genome-wide identification and expression analysis of beta-galactosidase family members in sweetpotato [*Ipomoea batatas* (L.) Lam.]. *BMC Genomics*. Available from <https://doi.org/10.21203/rs.3.rs-32133/v1>.
- Liu, J., Gao, M., Lv, M., & Cao, J. (2013). Structure, evolution, and expression of the β -galactosidase gene family in *Brassica campestris* ssp. *Chinensis*. *Plant Molecular Biology Reports*, 31, 1249–1260.
- Luthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, 356, 83–85.
- Masahiro Eda, M., Matsumoto, T., Ishimaru, M., & Tada, T. (2016). Structural and functional analysis of tomato β -galactosidase 4: Insight into the substrate specificity of the fruit softening-related enzyme. *The Plant Journal*, 86, 300–307.
- Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: A powerful approach for structure-based drug discovery. *Current Computer Aided Drug Design*, 7(2), 146–157.
- Moctezuma, E., Smith, D. L., & Gross, K. C. (2003). Antisense suppression of a β -galactosidase gene (TBG6) in tomato increases fruit cracking. *Journal of Experimental Botany*, 54(390), 2025–2033.
- Ng, J. K., Schroder, R., Brummell, D. A., Sutherland, P. W., Hallett, I. C., Smith, B. G., et al. (2015). Lower cell wall pectin solubilisation and galactose loss during early fruit development in apple (*Malus domestica*) cultivar ‘Scifresh’ are associated with slower softening rate. *Journal of Plant Physiology*, 176, 129–137.
- Ng, J. K., Schroder, R., Sutherland, P. W., Hallett, I. C., Hall, M. I., Prakash, R., et al. (2013). Cell wall structures leading to cultivar differences in softening rates develop early during apple (*Malus domestica*) fruit growth. *BMC Plant Biology*, 13, 183.
- Ohto, U., Usui, K., Ochi, T., Yuki, K., Satow, Y., & Shimizu, T. (2012). Crystal structure of human β -Galactosidase: Structural basis of gm1 gangliosidosis and morquioB diseases. *Journal of Biological Chemistry*, 287, 1801–1812.

- Ozeki, Y., Yokota, Y., Kato, K. H., Titani, K., & Matsui, T. (1995). Developmental expression of D-galactoside-binding lectin in sea urchin (*Anthocidaris crassispina*) eggs. *Experimental Cell Research*, *216*, 318–324.
- Petersen, T. N., Brunak, S., Heijne, G. V., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*, 785–786.
- Pressey, R. (1983). β -Galactosidases in ripening tomatoes. *Plant Physiology*, *71*, 132–135.
- Ranwala, A. P., Suematsu, C., & Masuda, H. (1992). The role of β -galactosidases in the modification of cell wall components during muskmelon fruit ripening. *Plant Physiology*, *100*, 1318–1325.
- Rojas, A. L., Nagem, R. A. P., Neustroev, K. N., Arand, M., Adamska, M., Eneyskaya, E. V., ... Polikarpov, I. (2004). Crystal structures of β -galactosidase from *Penicillium* sp. and its complex with galactose. *Journal of Molecular Biology*, *343*, 1281–1292.
- Ross, G. S., Redgwell, R. J., & MacRae, E. A. (1993). Kiwifruit β -galactosidase: Isolation and activity against specific fruit cell-wall polysaccharides. *Planta*, *189*, 499–506.
- Ross, G. S., Wagrzyn, T., MacRae, E. A., & Redgwell, R. J. (1994). Apple β -galactosidase: Activity against cell wall polysaccharides and characterization of a related cDNA clone. *Plant Physiology*, *106*, 521–528.
- Sethi, A., Joshi, K., Sasikala, K., & Alvala, M. (2019). In V. Gaitonde, P. Karmakar, & A. Trivedi (Eds.), *Molecular docking in modern drug discovery: Principles and recent applications, drug discovery and development – New advances*. IntechOpen. Available from <https://doi.org/10.5772/intechopen.85991>.
- Smith, D. L., Abbott, J. A., & Gross, K. C. (2002). Down-regulation of tomato β -galactosidase 4 results in decreased fruit softening. *Plant Physiology*, *129*, 1755–1762.
- Smith, D. L., & Gross, K. C. (2000). A family of at least seven β -galactosidase genes is expressed during tomato fruit development. *Plant Physiology*, *123*, 1173–1183.
- Szymanski, C. M., & Wren, B. W. (2005). Protein glycosylation in bacterial mucosal pathogens. *Nature Review Microbiology*, *3*, 225–237.
- Tanthanuch, W., Chantarangsee, M., Maneesan, J., & Ketudat-Cairns, J. (2008). Genomic and expression analysis of glycosyl hydrolase family 35 genes from rice (*Oryza sativa* L.). *BMC Plant Biology*, *8*, 84.
- Trainotti, L., Spinello, R., Piovan, A., Spolaore, S., & Casadoro, G. (2001). β -Galactosidases with a lectin-like domain are expressed in strawberry. *Journal of Experimental Botany*, *52*, 1635–1645.
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., ... Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*, W296–W303.
- Xu, D., & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal*, *101*, 2525–2534.
- Yang, H., Liu, J., Dang, M., Zhang, B., Li, H., Meng, R., ... Zhao, Z. (2018). Analysis of β -galactosidase during fruit development and ripening in two different texture types of apple cultivars. *Frontiers*, *9*, 539.

Next generation genomics: toward decoding domestication history of crops

Anjan Hazra and Sauren Das

Agricultural and Ecological Research Unit, Indian Statistical Institute, Kolkata, India

13.1 Introduction

Concurrent innovations in genomics have benefitted the crop domestication studies to a great extent. Comparative genomics comprising large scale population data of some existing crop varieties alongside their wild relatives may pave the way into the domestication history of species (Cao et al., 2014; Hufford et al., 2012). Certain archeological and population genetic data support that the initial domestication stages in Southwest Asia was a prolonged process (Fuller et al., 2014; Purugganan & Fuller, 2011) instead of a rapid evolution theory of cultivated plants which was presumed earlier (Abbo, Lev-Yadun, & Gopher, 2010; Hillman & Davies, 1990; Innan & Kim, 2004).

Genetic studies of crop domestication have been carried out since past decades through both top-down and bottom-up approaches. The classic top-down approaches are performed by means of analyzing the target phenotypic traits between wild and domesticated taxa and thereafter identifying the genetic variations that are correlated with the phenotypic traits (Doebley & Stec, 1991; Kantar, Nashoba, Anderson, Blackman, & Rieseberg, 2017; Paterson et al., 1988; Ross-Ibarra, Morrell, & Gaut, 2007; Sax, 1923). On the other hand, bottom-up approaches include examining the genetic variation among genomes of corresponding taxa, thereby dissection of domestication related evolutionary signals leading to integrate it with the domesticated phenotypes (Kantar et al., 2017; Ross-Ibarra et al., 2007). Recent advent of high-throughput sequencing technologies enabled to compare the whole genomes of representative individuals from domesticated taxa with their wild relatives (Hufford et al., 2012; Li et al., 2013; Wang et al., 2019; Yang et al., 2012; Zeng et al., 2019). Genome-wide genetic markers enhance our understanding regarding the global and local evolutionary signals evident throughout the genome (Diao & Chen, 2012). They can distinguish the signals of selection during domestication (Vitti, Grossman, & Sabeti, 2013) from other definite signals related to demographic needs (Guerra-García & Piñero, 2017; Meyer & Purugganan, 2013) (Fig. 13.1).

13.2 Whole genome sequencing

Whole-genome sequence of a species is the key resource in modern domestication studies. It provides the reference genome of the crop, a prerequisite of all downstream genomic analyses. Whole-genome sequencing and assembly entirely dependent on the various next generation sequencing modules. These high-throughput sequencing technologies include Illumina (Sun et al., 2017), PacBio (Badouin et al., 2017; VanBuren et al., 2018), Oxford Nanopore (Belser et al., 2018) or a combination of these (Bickhart et al., 2017; Zhou et al., 2019). Prior to the sequencing project initiation, the haploid genome size and the ploidy of the organism are determined to estimate the assembly strategy and sequencing expenditure (Sims, Sudbery, Ilott, Heger, & Ponting, 2014). Since eukaryotic genomes consist of a major amount of repetitive elements, the sequencing libraries are being prepared with large insert sizes denoted as mate-pair libraries (Barrera-Redondo, Piñero, & Eguiarte, 2020). Alternatively, long-read sequencing technologies such as PacBio or Oxford Nanopore (Levy & Myers, 2016; Sohn & Nam, 2018) can also be employed. Following genome sequencing and assembling, it must be properly annotated. It requires the transcriptome data from the same species, as well as the homology evidenced from other curated genomes and *ab initio* predictions based on the underlying structure of genes (Yandell & Ence, 2012).

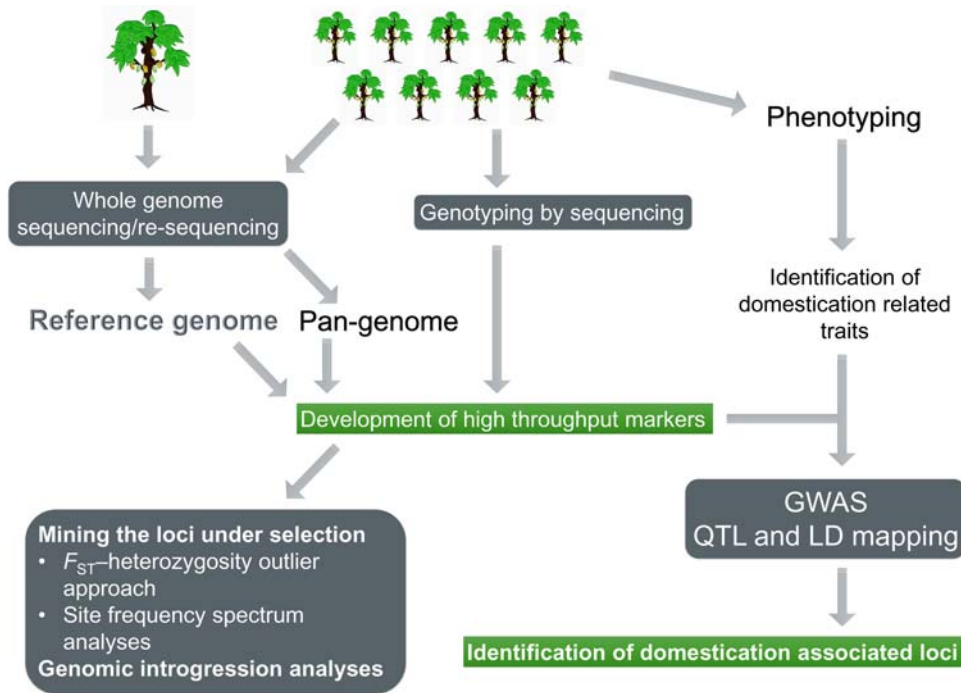


FIGURE 13.1 Workflow showing current approaches of domestication studies using genomic tools.

In recent years, the sequencing cost per nucleotide in all the aforementioned technologies has become depleted and thus the whole-genome assembly projects are affordable to more researcher groups (Consortium, 2012; Schnable et al., 2009) and become popularized widely (Muir et al., 2016). Bottlenecks still exist to small research groups in handling the massive amounts of genome data within limited availability of computational resources towards storage and analysis (Barrera-Redondo et al., 2020; Muir et al., 2016) (Table 13.1). Therefore availability of reference genome is pivotal for genomic studies towards domestication of a crop (Barrera-Redondo et al., 2020). Since domesticated taxa are mostly having economic importance, it becomes easier to persuade the funding agencies to support the genome projects. Thus, reference genomes for most of the domesticated species are now available which might be employed for crop improvement programs as well (Ellegren, 2014).

13.3 Alternative genome scale approaches

For organisms with very large genomes or in the absence of their reference genome, sequencing arbitrary and/or desired portions of the genome or sequencing the transcriptionally active portions of the genome might be useful in detecting the genetic variations (Mastretta-Yanes et al., 2015; Schreiber, Stein, & Mascher, 2018). In fact, the reduced sequencing cost per sample allows for a large sample size alongside a high sequencing depth leading to higher accuracy of the observed genetic variation (De Mita et al., 2013; Lotterhos & Whitlock, 2015; Schreiber et al., 2018). Although this method covers a fraction of the whole genome, still it is incredibly sufficient to infer the population genetic statistics, to detect signatures of selective sweeps and GWAS for domestication related traits (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Schreiber et al., 2018).

Restriction site-associated DNA sequencing (RAD-seq) or the updated double digest Restriction site-associated DNA sequencing (ddRAD-seq) has recently become a popular, convenient, cost-effective genotyping by sequencing approach for population genomic studies (Davey & Blaxter, 2010). The technique involves one or two restriction enzymes to digest the genomic DNA and subsequently sequence the regions adjacent to the restriction sites scattered across the genome (Davey & Blaxter, 2010). The resultant sequence data can either be mapped against a reference genome or instead it can be assembled *de novo* (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; Hazra, Kumar, Sengupta, & Das, 2021; Mastretta-Yanes et al., 2015) which makes it a versatile technique for species with limited genomic resources (Barrera-Redondo et al., 2020). However, the reference-based approach is highly recommended to avoid the possible downstream errors in the estimation of site frequency spectrum (Shafer et al., 2017).

TABLE 13.1 Comparison of sequencing platforms with the features and performances. (Kulski, 2016)

Generation	Platform	Company	Read length per run (bp)	Reads per run	Time	Cost per 10 ⁶ bases	Raw error rate (%)	Platform cost (USD approx.)	Chemistry
First generation	Sanger	Life Technologies	800	1	2 h	2400	0.3	95,000	Dideoxy terminator
Second generation	454 GS FLX +	Roche	700	1 × 10 ⁶	24/48 h	10	1	500,000	Pyrosequencing
	HiSeq	Illumina	2 × 150	5 × 10 ⁹	27/240 h	0.1	0.8	750,000	Reversible terminators
	MiSeq	Illumina	2 × 300	3 × 10 ⁸	27 h	0.13	0.8	125,000	Reversible terminators
	SOLiD	Life Technologies	50	1 × 10 ⁹	14 days	0.13	0.01	350,000	Ligation
	RetrovLOCITY	BGI	50	1 × 10 ⁹	14 days	0.01	0.01	12 × 10 ⁶	Nanoball/ligation
	Ion Proton	Life Technologies	200	6 × 10 ⁷	2–5 h	1	1.7	215,000	Proton detection
	Ion PGM	Life Technologies	200	5 × 10 ⁶	2–5 h	1	1.7	80,000	Proton detection
Third generation	SMRT	PacBio	>10,000	1 × 10 ⁶	1–2 h	2	12.9	750,000	Real-time SMS
	Heliscope	Helicos	35	7 × 10 ⁹	8 days	0.01	0.2	1.35 × 10 ⁶	Real-time SMS
	Nanopore	Oxford Nanopore Technologies	>5000	6 × 10 ⁴	48/72 h	<1	34	1000	Real-time SMS

Source: Data from Kulski J.K. Next-generation sequencing—an overview of the history, tools, and “omic” applications. Next generation sequencing-advances, applications and challenges. 2016:3-60.

Transcriptome sequencing (RNA-seq) is a popular approach to obtain population-level variations in the transcriptionally active part of the genomes (De Wit et al., 2012). Exome sequencing is another suitable alternative targeting the protein-coding regions of the genome (Kaur & Gaikwad, 2017; Warr et al., 2015) and the demographic history and selective sweeps can be estimated using this method (Pankin, Altmüller, Becker, & von Korff, 2018). However the limitations of exome/transcriptome sequencing method are to the selective signals those occur restriction to the transcriptionally active/ protein coding part of the genome, whereas many of the domestication associated genetic changes are located within *cis*- and *trans*-regulatory elements, noncoding RNAs (Swinnen, Goossens, & Pauwels, 2016).

13.4 Emergence of pan-genomics

Structural variants at the genome level (copy-number variation, presence/absence of genomic regions, inversions, transversions, translocations) are very common within organisms (Khan et al., 2020). Moreover, the structural variants including copy-number disparity play a pivotal role underlying functional variation of genes as well as the emergence of diversification and domestication traits among the crop varieties (Hazra, Dasgupta, Sengupta, & Das, 2019; Lye & Purugganan, 2019). For example, at least one third of the known domestication related loci found to be structural variants in grapevine individuals (Zhou et al., 2019). Therefore a single reference genome never represents the full repertoire of all strains within the species (Munir et al., 2020; Zhao et al., 2018). This led to the generation of the concept “Pan-genome” that first reported in microbiology (Tettelin et al., 2005), and later on into the work-domain of plants and animals as well (Golicz, Batley, & Edwards, 2016). The copy number variations, presence/absence variations (PAVs), and SNPs altogether can serve as the basis of adaptation of the species within a specific environment or any selective regimes (Lye & Purugganan, 2019).

Once pan-genome data is available for an organism, it can be utilized in analyzing the structural variants in populations leading to reveal novel loci involved in the development of domestication-related traits (Li et al., 2014; Zhao et al., 2018). Pan-genomes have become available for several plant species such as maize (Brohammer, Kono, & Hirsch, 2018), wheat (Montenegro et al., 2017), *Brassica* (Golicz et al., 2016; Hurgobin et al., 2018). Pan-genomes are also implemented in domestication studies for several crops, that is, soybean (Li et al., 2014), rice (Zhao et al., 2018), sunflower (Hübner et al., 2019) and tomato (Gao et al., 2019).

13.5 Methodologies in domestication genomics

On the availability of the genomic resources of a crop and its wild relatives, the subsequent necessary tasks are to analyze the data toward inference of the population structure, genetic variations, selection pattern and identification of important loci that are pointing towards the shape of domestication process. Concurrent genome editing tools would conveniently assist the validation of marker trait association. A comprehensive list of widely used relevant genomic tools is being provided in Table 13.2.

13.6 Case studies on next-generation sequencing-assisted inference of domestication history

13.6.1 Rice

Domestication and history of origin are more or less complex in case of cultivated rice (*Oryza sativa*) varieties. The japonica cultivars consist of 1 of 3 types of chloroplast genome among which at least one was derived from *O. rufipogon*, with the same cp genome type. A polyphyletic origin of cultivated Asian rice from 4 different lineages in the *O. rufipogon* and *O. nivara* complex was reported (Kawakami et al., 2007). Cheng et al. (2019) focused on the domestication of Asian rice (*indica* and *japonica*) through the selection characteristics in the chloroplast genome that occurred in different Asian rice during the domestication process. Diversity and phylogenetic analyses revealed, *Oryza sativa* L. ssp. *japonica* possess slightly less diversity (π) than *Oryza sativa* L. ssp. *indica* and wild rice. The results indicated that Asian rice had been domesticated at least twice. Civián et al. (2019) concluded that the aromatic rice arose independently in the Indian subcontinent and thereby the japonica population arrived here some 4,000 years ago. Later on, the japonica varieties accomplished aroma traits through hybridization with wild rice along the foothills of the Himalayas.

TABLE 13.2 Various methods and tools used in plant domestication genomics.

Methods	Test/analyses	Principle	Examples of implementation
Population genetics	<i>Genetic Diversity in Populations</i> —allele frequencies, heterozygosity, nucleotide diversity, number of segregating sites and private alleles	Reveal the level of genetic erosion in domesticated plants compared to the ancestral wild population, caused by bottlenecks, selective sweeps and inbreeding (Gepts, 2014; Groeneveld et al., 2010)	Most of the following studies
	<i>Population Structure</i> —Parametric and Non-parametric methods (principal component analyses, discriminant analyses of principal components and K-means clustering)	Reveal the influence of historical events that shaped the genetic diversity of the organisms (Linck & Battey, 2019)	
	<i>Temporal changes of effective population size</i>	Understanding the evolutionary aspects of domestication concerning natural and artificial selection (Allaby, Ware, & Kistler, 2019; Chen et al., 2018)	
Ancient gene flow and local ancestry	Graph-based	The relationships between populations as a bifurcating tree, represents ancient gene flow that contributed to modern genetic variation (Pickrell & Pritchard, 2012)	Pearl millet (Burgarella et al., 2018)
	ABBA-BABA test or D-statistic	Evaluates the allelic patterns of three taxa and compares them to an outgroup to identify genomic regions with an excess of shared derived variants that are not concordant to the species tree (i.e., ABBA-BABA patterns), which suggest introgression events (Durand, Patterson, Reich, & Slatkin, 2011)	Evolution of C ₄ photosynthesis (Olofsson et al., 2016)
Demographic Simulations	Approximate Bayesian computation (ABC) method	Compares the summary statistics of simulations against the observed data to accept or reject certain demographic hypotheses (Cornuet et al., 2014; Gerbault et al., 2014)	Scarlet runner bean (Guerra-García & Piñero, 2017)
Identifying genes under selection	F _{ST} Outlier Tests	Detect signals of selective sweeps between populations of wild and domesticated taxa (Gepts, 2014)	Evolution of fruit quality in Apple (Khan, Olsen, Sovero, Kushad, & Korban, 2014) and oil properties in sunflower (Baute, Kane, Grassa, Lai, & Rieseberg, 2015)
	<i>Site Frequency Spectrum Based Tests</i>		
	Tajima's D	It is sensitive to changes in low-frequency variants, making it particularly useful to detect selective sweeps before and after the selected locus reaches fixation, although low-frequency variants can also be observed in loci under purifying selection (Tajima, 1989; Zeng, Fu, Shi, & Wu, 2006).	Sunflower (Baute et al., 2015)

(Continued)

TABLE 13.2 (Continued)

Methods	Test/analyses	Principle	Examples of implementation
	Fay and Wu's H	It is sensitive to changes in high frequency variants, which are only altered by positive selection, making it very useful when used alongside Tajima's D (Fay & Wu, 2000)	Peach and almond (Velasco, Hough, Aradhya, & Ross-Ibarra, 2016)
	Zeng et al.'s E (Zeng et al., 2006)	It is sensitive to both low and high frequency variants, making it particularly powerful to detect selective sweeps before or after the selected locus reached fixation, also needing an outgroup in order to differentiate derived alleles (from ancestral alleles).	
	Reduction of diversity (ROD)	It compares local π values of domesticated taxa against the local π values of its wild relatives, using sliding windows alongside the genome (Guo et al., 2013; Schmutz et al., 2014)	Rice (Huang et al., 2012), Watermelon (Guo et al., 2013), Cucumber (Qi et al., 2013), Common bean (Schmutz et al., 2014), and Chickpea (Varshney et al., 2019)
	<i>Linkage Disequilibrium (LD) Based Methods</i>		
	EHH statistics, LD decay (LDD) test	Given that selective sweeps remove the variation in regions adjacent to the locus under selection, they can form haplotype blocks that extend in strong LD compared to other haplotypes in the same locus (Sabeti et al., 2002; Vitti et al., 2013)	Potato (Vos et al., 2017)
	XP-CLR test (Chen, Patterson, & Reich, 2010)	Implement multiple signatures to detect selective sweeps	Maize (Hufford et al., 2012)
	μ statistic (Alachiotis & Pavlidis, 2018)	implement multiple signatures to detect selective sweeps	African rice (Ndjiondjop et al., 2019)
Genome wide association studies		Unravel the genetic variants underlying the domestication traits	Chickpea (Varshney et al., 2019), Peach (Cao et al., 2019), Rice (Zheng et al., 2019)
Paleogenomics		Extraction and Sequencing of Ancient DNA	Maize (Ramos-Madrigal et al., 2016; Vallebuena-Estrada et al., 2016), Grapevine (Wales et al., 2016), Barley (Mascher et al., 2016), Sunflower (Wales et al., 2019)
Transcriptome analyses	Differential Expression Analyses	Involvement of the genes toward phenotypic differences associated to domestication	Maize and teosinte (Swanson-Wagner et al., 2012), Tomato (Koenig et al., 2013), Pea (Hradilová et al., 2017), Common bean (Singh, Zhao, & Vallejos, 2018), and carrot (Machaj, Bostan, Macko-Podgórní, Iorizzo, & Grzebelus, 2018)

(Continued)

TABLE 13.2 (Continued)

Methods	Test/analyses	Principle	Examples of implementation
Epigenomic studies	Epigenome-wide association studies (EWAS), Single methylation polymorphisms (SMPs)	Epigenetic marks associated to transcriptional gene silencing (He, Chen, & Zhu, 2011)	Cotton (Song, Zhang, Stelly, & Chen, 2017), Soybean (Shen et al., 2018)
Experimental validation of domestication loci	<i>Genome editing tools</i> – 1. Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR) system alongside the CRISPR associated protein 9 (Cas9), commonly known as CRISPR/Cas9 (Cong et al., 2013). 2. Transcription Activator-Like Effector Nuclease (TALEN) technology (Zhang, Zhang, Lang, Botella, & Zhu, 2017)	In vitro techniques by knock-out, knock-down or knock-in experiments to validate the involvement of the predicted genes or loci in the observed phenotypes (Zhang et al., 2017).	

13.6.2 Citrus

Complex domestication and origin history was also dissected through genomic approaches in case of Citrus (Wu et al., 2018). This study included 60 accessions representing diverse citrus germplasms and ten natural citrus species for genomic, phylogenetic and biogeographic analyses. Accordingly, the citrus fruit proposed to be originated in the southeast foothills of the Himalayas, thereafter underwent a sudden speciation during Miocene, segregated through dispersal from southeast Asia to Australasia, and later adapted to both these diverse climates (Wu et al., 2018). Subsequent analyses of hybrids and admixed genomes therein resulted into the genealogy of major commercial cultivars of citrus. An extensive network of relatedness among mandarins and sweet orange indicated domestication signals of these groups. Moreover, admixture of pummelo among these mandarins and their correlation with the fruit size and acidity suggested a reasonable background of pummelo introgression while the selection of palatable mandarins are concerned.

13.6.3 Peanut

The origin of peanut (*Arachis hypogaea*) has been attempted to trace through several genomic studies during recent times (Bertioli et al., 2019; Chen et al., 2019; Zhuang et al., 2019). Multiple evidence of peanut genome analyses indicated a recent (<10,000 years ago) origin of this important crop that was aided by demographic activities (Bertioli et al., 2016, 2019). Furthermore, Bertioli, Abernathy, Seijo, Clevenger, and Cannon (2020) have assessed the different models for peanut's origin which strongly support the identities of the two ancestral species as *A. duranensis* and *A. ipaensis*. According to them, *A. ipaensis* was moved by humans in ancient times which may be a direct descendant of the population that gave rise to cultivated peanuts in course of polyploidization events <10,000 years ago (Bertioli et al., 2020). However, based on K_s values for collinear gene pairs and the expansion time of transposable elements, Zhuang et al. (2019) suggested that peanut polyploidization occurred ~450,000 years ago and also the previously claimed *A. duranensis* was not a direct descendent of the ancestor of the A-subgenome donor. Supporting this, Zhuang et al. (2020) further opined that cultivated peanut originated sometimes 0.40 million years ago by natural interspecific hybridization followed by polyploidization and the tetraploid peanuts possibly have experienced genetic exchanges with diploids more recently, representing its evolutionary complexity.

13.6.4 Olive

The origins and the genetic background of the domestication related phenotypic changes are still debatable since the last report (Gros-Balthazard et al., 2019). RNA-sequencing data of 68 wild and cultivated olive trees were considered to identify the differentially expressed genes and genetic signatures for selection exercise during domestication process (Gros-Balthazard et al., 2019). This breakthrough report suggests a major domestication event in the eastern part of the

Mediterranean basin followed by dispersion towards the West and subsequent admixture with western wild olives. The transcriptome wide investigation uncovered the domestication traits of olive mainly arose through changes in gene expression which is consistent with its evolutionary history and life history traits.

13.6.5 Tea

The highest tea producing country of the world, China, harbors abundant tea germplasms and has long been considered the cradle of origin of tea (Hasimoto, 2001; Xia et al., 2020). Nevertheless, due to undiscovered wild ancestors of tea plants in China, the domestication story of tea plants is still in an enigma (Hazra et al., 2021; Xia et al., 2020). Meegahakumbura et al. (2018) claimed that China type tea and Assam type tea first diverged 22,000 years ago and subsequently split into the Chinese Assam type tea and Indian Assam type tea lineages at about 2770 years ago. However, the statements are unilateral and thus, controversy still alive due to the sampling biases and lacking sufficient evidence with molecular marker proofs (Xia et al., 2020). Since most of the existing cultivated varieties of tea are the result of constant breeding and hybridizations occurred in natural populations, so artificial domestication may have had little impact on the variation in genome sequences. Recent availability of the tea plant genome provides reasonable opportunity in solving this disparity. However, following measures should be taken into account for resolving tea origin and domestication debate: (1) collection of worldwide representative of tea plant samples; (2) investigation of the population structure and putative wild ancestor of tea plants; (3) estimation of population diversity employing genome-wide SNP markers; (4) identification of the candidate genomic regions selected during domestication; and (5) functional investigation of the domestication related genes (Xia et al., 2020). In a recent study, genomic re-sequencing was carried out in 139 global tea accessions for population genomics and evolutionary studies in tea (Wang et al., 2020). Phylogenetic analyses in this investigation revealed that the selection for favorable disease resistance and flavor traits during domestication has been predominant in *C. sinensis* var. *sinensis* populations than that of *C. sinensis* var. *assamica* populations.

References

- Abbo, S., Lev-Yadun, S., & Gopher, A. (2010). Agricultural origins: Centers and noncenters; a near eastern reappraisal. *Critical Reviews in Plant Science*, 29(5), 317–328.
- Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications biology*, 1(1), 1–11.
- Allaby, R. G., Ware, R. L., & Kistler, L. (2019). A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evolutionary Applications*, 12(1), 29–37.
- Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics*, 17(2), 81.
- Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, 546(7656), 148–152.
- Barrera-Redondo, J., Piñero, D., & Eguiarte, L. E. (2020). Genomic, transcriptomic and epigenomic tools to study the domestication of plants and animals: a field guide for beginners. *Frontiers in Genetics*, 11, 742.
- Baute, G. J., Kane, N. C., Grassa, C. J., Lai, Z., & Riieseberg, L. H. (2015). Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytologist*, 206(2), 830–838.
- Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., et al. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature plants*, 4(11), 879–887.
- Bertioli, D. J., Abernathy, B., Seijo, G., Clevenger, J., & Cannon, S. B. (2020). Evaluating two different models of peanut's origin. *Nature Genetics*, 1–3.
- Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics*, 48(4), 438–446.
- Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., et al. (2019). The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nature Genetics*, 51(5), 877–884.
- Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., et al. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, 49(4), 643–650.
- Brohammer, A. B., Kono, T. J., & Hirsch, C. N. (2018). The maize pan-genome. *The maize genome*, 13–29.
- Burgarella, C., Cubry, P., Kane, N. A., Varshney, R. K., Mariac, C., Liu, X., et al. (2018). A western Sahara centre of domestication inferred from pearl millet genomes. *Nature Ecology & Evolution*, 2(9), 1377–1380.
- Cao, K., Li, Y., Deng, C. H., Gardiner, S. E., Zhu, G., Fang, W., et al. (2019). Comparative population genomics identified genomic regions and candidate genes associated with fruit domestication traits in peach. *Plant Biotechnology Journal*, 17(10), 1954–1970.

- Cao, K., Zheng, Z., Wang, L., Liu, X., Zhu, G., Fang, W., et al. (2014). Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biology*, 15(7), 1–15.
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140.
- Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research*, 20(3), 393–402.
- Chen, J., Ni, P., Li, X., Han, J., Jakovlić, I., Zhang, C., et al. (2018). Population size may shape the accumulation of functional mutations following domestication. *BMC Evolutionary Biology*, 18(1), 1–10.
- Chen, X., Lu, Q., Liu, H., Zhang, J., Hong, Y., Lan, H., et al. (2019). Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Molecular plant*, 12(7), 920–934.
- Cheng, L., Nam, J., Chu, S.-H., Rungnana, P., Min, M.-h, Cao, Y., et al. (2019). Signatures of differential selection in chloroplast genome between japonica and indica. *Rice*, 12(1), 65.
- Civáň, P., Ali, S., Batista-Navarro, R., Drosou, K., Ihejieta, C., Chakraborty, D., et al. (2019). Origin of the aromatic group of cultivated rice (*Oryza sativa* L.) traced to the Indian subcontinent. *Genome biology and evolution*, 11(3), 832–843.
- Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, 339(6121), 819–823.
- Consortium, T. G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400), 635.
- Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., et al. (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics (Oxford, England)*, 30(8), 1187–1189.
- Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in functional genomics*, 9(5–6), 416–423.
- De Mita, S., Thuillet, A. C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, 22(5), 1383–1399.
- De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., et al. (2012). The simple fool's guide to population genomics via RNA-Seq: An introduction to high-throughput sequencing data analysis. *Molecular ecology resources*, 12(6), 1058–1067.
- Diao, L., & Chen, K. C. (2012). Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. *Genetics*, 192(4), 1503–1511.
- Doebley, J., & Stec, A. (1991). Genetic analysis of the morphological differences between maize and teosinte. *Genetics*, 129(1), 285–295.
- Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8), 2239–2252.
- Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, 29(1), 51–63.
- Fay, J. C., & Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3), 1405–1413.
- Fuller, D. Q., Denham, T., Arroyo-Kalin, M., Lucas, L., Stevens, C. J., Qin, L., et al. (2014). Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. *Proceedings of the National Academy of Sciences*, 111(17), 6147–6152.
- Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, 51(6), 1044–1051.
- Gepts, P. (2014). The contribution of genetic and genomic approaches to plant domestication studies. *Current Opinion in Plant Biology*, 18, 51–59.
- Gerbault, P., Allaby, R. G., Boivin, N., Ruzdinski, A., Grimaldi, I. M., Pires, J. C., et al. (2014). Storytelling and story testing in domestication. *Proceedings of the National Academy of Sciences*, 111(17), 6159–6164.
- Golicz, A. A., Batley, J., & Edwards, D. (2016). Towards plant pangenomics. *Plant Biotechnology Journal*, 14(4), 1099–1105.
- Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., et al. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, 7(1), 1–8.
- Groeneveld, L., Lenstra, J., Eding, H., Toro, M., Scherf, B., Pilling, D., et al. (2010). Genetic diversity in farm animals—a review. *Animal Genetics*, 41, 6–31.
- Gros-Balthazard, M., Besnard, G., Sarah, G., Holtz, Y., Leclercq, J., Santoni, S., et al. (2019). Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. *The Plant Journal*, 100(1), 143–157.
- Guerra-García, A., & Piñero, D. (2017). Current approaches and methods in plant domestication studies. *Botanical Sciences*, 95(3), 345–362.
- Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., et al. (2013). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nature Genetics*, 45(1), 51–58.
- Hasimoto M., editor The origin of the tea plant. Proceedings of 2001 International Conference on O–Cha (Tea) Culture and Science (Session II), October; 2001.
- Hazra, A., Dasgupta, N., Sengupta, C., & Das, S. (2019). MIPS: Functional dynamics in evolutionary pathways of plant kingdom. *Genomics*, 111(6), 1929–1945.
- Hazra, A., Kumar, R., Sengupta, C., & Das, S. (2021). Genome-wide SNP discovery from Darjeeling tea cultivars—their functional impacts and application toward population structure and trait associations. *Genomics*, 113(1), 66–78.
- Hazra, A., Mahadani, P., Das, S., Bhattacharya, S., Kumar, R., Sengupta, C., et al. (2021). Insight to the ancestral relations and varietal diversity of Indian tea [*Camellia sinensis* (L.) Kuntze] through plastid and nuclear phylogenetic markers. *Genetic Resources and Crop Evolution*, 68, 773–783.
- He, X.-J., Chen, T., & Zhu, J.-K. (2011). Regulation and function of DNA methylation in plants and animals. *Cell Research*, 21(3), 442–465.
- Hillman, G. C., & Davies, M. S. (1990). 6. Domestication rates in wild-type wheats and barley under primitive cultivation. *Biological Journal of the Linnean Society*, 39(1), 39–78.

- Hradilová, I., Trněný, O., Váľková, M., Cechová, M., Janská, A., Prokešová, L., et al. (2017). A combined comparative transcriptomic, metabolomic, and anatomical analyses of two key domestication traits: pod dehiscence and seed dormancy in pea (*Pisum* sp.). *Frontiers in Plant Science*, 8, 542.
- Huang, X., Kurata, N., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 490(7421), 497–501.
- Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., et al. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature Plants*, 5(1), 54–62.
- Hufford, M. B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nature Genetics*, 44(7), 808–811.
- Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C. K. K., Tirnaz, S., Dolatabadian, A., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal*, 16(7), 1265–1274.
- Innan, H., & Kim, Y. (2004). Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences*, 101(29), 10667–10672.
- Kantar, M. B., Nashoba, A. R., Anderson, J. E., Blackman, B. K., & Rieseberg, L. H. (2017). The genetics and genomics of plant domestication. *Bioscience*, 67(11), 971–982.
- Kaur, P., & Gaikwad, K. (2017). From genomes to GENE-omes: exome sequencing concept and applications in crop improvement. *Frontiers in Plant Science*, 8, 2164.
- Kawakami, S.-i., Ebana, K., Nishikawa, T., Sato, Y.-i., Vaughan, D. A., & Kadowaki, K.-i (2007). Genetic variation in the chloroplast genome suggests multiple domestication of cultivated Asian rice (*Oryza sativa* L.). *Genome/National Research Council Canada; Genome/Conseil National de Recherches Canada*, 50(2), 180–187.
- Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., & Varshney, R. K. (2020). Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends in Plant Science*, 25(2), 148–158.
- Khan, M. A., Olsen, K. M., Sovero, V., Kushad, M. M., & Korban, S. S. (2014). Fruit quality traits have played critical roles in domestication of the apple. *The Plant Genome*, 7(3), plantgenome2014.04.0018.
- Koenig, D., Jiménez-Gómez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., et al. (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proceedings of the National Academy of Sciences*, 110(28), E2655–E2662.
- Kulski, J. K. (2016). Next-generation sequencing—An overview of the history, tools, and “omic” applications. *Next generation sequencing-advances, applications and challenges*, 3–60.
- Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, 17, 95–115.
- Li, Y., vonHoldt, B. M., Reynolds, A., Boyko, A. R., Wayne, R. K., Wu, D.-D., et al. (2013). Artificial selection on brain-expressed genes during the domestication of dog. *Molecular Biology and Evolution*, 30(8), 1867–1876.
- Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, 32(10), 1045–1052.
- Linck, E., & Battey, C. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, 19(3), 639–647.
- Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, 24(5), 1031–1046.
- Lye, Z. N., & Purugganan, M. D. (2019). Copy number variation in domestication. *Trends in Plant Science*, 24(4), 352–365.
- Machaj, G., Bostan, H., Macko-Podgórn, A., Iorizzo, M., & Grzebelus, D. (2018). Comparative transcriptomics of root development in wild and cultivated carrots. *Genes*, 9(9), 431.
- Mascher, M., Schuenemann, V. J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., et al. (2016). Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nature Genetics*, 48(9), 1089–1093.
- Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular ecology resources*, 15(1), 28–41.
- Meegahakumbura, M. K., Wambulwa, M. C., Li, M.-M., Thapa, K. K., Sun, Y.-S., Möller, M., et al. (2018). Domestication origin and breeding history of the tea plant (*Camellia sinensis*) in China and India based on nuclear microsatellites and cpDNA sequence data. *Frontiers in Plant Science*, 8, 2270.
- Meyer, R. S., & Purugganan, M. D. (2013). Evolution of crop species: Genetics of domestication and diversification. *Nature Reviews. Genetics*, 14(12), 840–852.
- Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K. K., et al. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5), 1007–1013.
- Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., et al. (2016). The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biology*, 17(1), 1–9.
- Munir, F., Saba, N. U., Arveen, M., Siddiq, A., Ahmad, J., & Amir, R. (2020). *Pan-genomics of plants and its applications. Pan-genomics: Applications, Challenges, and Future Prospects* (pp. 285–306). Elsevier.
- Ndjiondjop, M. N., Alachiotis, N., Pavlidis, P., Goungoulou, A., Kpeki, S. B., Zhao, D., et al. (2019). Comparisons of molecular diversity indices, selective sweeps and population structure of African rice with its wild progenitor and Asian rice. *Theoretical and Applied Genetics*, 132(4), 1145–1158.
- Olofsson, J. K., Bianconi, M., Besnard, G., Dunning, L. T., Lundgren, M. R., Holota, H., et al. (2016). Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology*, 25(24), 6107–6123.

- Pankin, A., Altmüller, J., Becker, C., & von Korff, M. (2018). Targeted resequencing reveals genomic signatures of barley domestication. *New Phytologist*, 218(3), 1247–1259.
- Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, 335(6192), 721–726.
- Pickrell, J., & Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, 1.
- Purugganan, M. D., & Fuller, D. Q. (2011). Archaeological data reveal slow rates of evolution during plant domestication. *Evolution: International Journal of Organic Evolution*, 65(1), 171–183.
- Qi, J., Liu, X., Shen, D., Miao, H., Xie, B., Li, X., et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nature Genetics*, 45(12), 1510.
- Ramos-Madrugal, J., Smith, B. D., Moreno-Mayar, J. V., Gopalakrishnan, S., Ross-Ibarra, J., Gilbert, M. T. P., et al. (2016). Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Current Biology*, 26(23), 3195–3201.
- Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences*, 104(suppl 1), 8641–8648.
- Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909), 832–837.
- Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*, 8(6), 552.
- Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, 46(7), 707–713.
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, 326(5956), 1112–1115.
- Schreiber, M., Stein, N., & Mascher, M. (2018). Genomic approaches for studying crop evolution. *Genome Biology*, 19(1), 1–15.
- Shafer, A. B., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., et al. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 8(8), 907–917.
- Shen, Y., Zhang, J., Liu, Y., Liu, S., Liu, Z., Duan, Z., et al. (2018). DNA methylation footprints during soybean domestication and improvement. *Genome Biology*, 19(1), 1–14.
- Sims, D., Sudbery, I., Illott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics*, 15(2), 121–132.
- Singh, J., Zhao, J., & Vallejos, C. E. (2018). Differential transcriptome patterns associated with early seedling development in a wild and a domesticated common bean (*Phaseolus vulgaris* L.) accession. *Plant Science*, 274, 153–162.
- Sohn, J.-i, & Nam, J.-W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, 19(1), 23–40.
- Song, Q., Zhang, T., Stelly, D. M., & Chen, Z. J. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biology*, 18(1), 1–14.
- Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., et al. (2017). Karyotype stability and unbiased fractionation in the paleo-allotetraploid Cucurbita genomes. *Molecular plant*, 10(10), 1293–1306.
- Swanson-Wagner, R., Briskine, R., Schaefer, R., Hufford, M. B., Ross-Ibarra, J., Myers, C. L., et al. (2012). Reshaping of the maize transcriptome by domestication. *Proceedings of the National Academy of Sciences*, 109(29), 11878–11883.
- Swinnen, G., Goossens, A., & Pauwels, L. (2016). Lessons from domestication: targeting cis-regulatory elements for crop improvement. *Trends in Plant Science*, 21(6), 506–515.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585–595.
- Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences*, 102(39), 13950–13955.
- Vallebuena-Estrada, M., Rodríguez-Arévalo, I., Rougon-Cardoso, A., González, J. M., Cook, A. G., Montiel, R., et al. (2016). The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proceedings of the National Academy of Sciences*, 113(49), 14151–14156.
- VanBuren, R., Wai, C. M., Colle, M., Wang, J., Sullivan, S., Bushakra, J. M., et al. (2018). A near complete, chromosome-scale assembly of the black raspberry (*Rubus occidentalis*) genome. *Gigascience*, 7(8), giy094.
- Varshney, R. K., Thudi, M., Roorkiwal, M., He, W., Upadhyaya, H. D., Yang, W., et al. (2019). Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nature Genetics*, 51(5), 857–864.
- Velasco, D., Hough, J., Aradhya, M., & Ross-Ibarra, J. (2016). Evolutionary genomics of peach and almond domestication. *G3: Genes, Genomes, Genetics*, 6(12), 3985–3993.
- Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, 47, 97–120.
- Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G., van Eck, H. J., & van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics*, 130(1), 123–135.
- Wales, N., Akman, M., Watson, R. H., Sánchez Barreiro, F., Smith, B. D., Gremillion, K. J., et al. (2019). Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3,000 years of sunflower domestication and improvement. *Evolutionary applications*, 12(1), 38–53.
- Wales, N., Madrugal, J. R., Cappellini, E., Baez, A. C., Castruita, J. A. S., Romero-Navarro, J. A., et al. (2016). The limits and potential of paleogenomic techniques for reconstructing grapevine domestication. *Journal of Archaeological Science*, 72, 57–70.

- Wang, W., Zhang, X., Zhou, X., Zhang, Y., La, Y., Zhang, Y., et al. (2019). Deep genome resequencing reveals artificial and natural selection for visual deterioration, plateau adaptability and high prolificacy in Chinese domestic sheep. *Frontiers in genetics, 10*, 300.
- Wang, X., Feng, H., Chang, Y., Ma, C., Wang, L., Hao, X., et al. (2020). Population sequencing enhances understanding of tea plant evolution. *Nature communications, 11*(1), 1–10.
- Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome sequencing: current and future perspectives. *G3: Genes, Genomes, Genetics., 5*(8), 1543–1550.
- Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C., et al. (2018). Genomics of the origin and evolution of Citrus. *Nature, 554*(7692), 311–316.
- Xia, E.-H., Tong, W., Wu, Q., Wei, S., Zhao, J., Zhang, Z.-Z., et al. (2020). Tea plant genomics: achievements, challenges and perspectives. *Horticulture Research, 7*(1), 1–19.
- Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics, 13*(5), 329–342.
- Yang, L., Koo, D. H., Li, Y., Zhang, X., Luan, F., Havey, M. J., et al. (2012). Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. *The Plant Journal, 71*(6), 895–906.
- Zeng, K., Fu, Y.-X., Shi, S., & Wu, C.-I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics, 174*(3), 1431–1439.
- Zeng, L., Tu, X.-L., Dai, H., Han, F.-M., Lu, B.-S., Wang, M.-S., et al. (2019). Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biology, 20*(1), 1–13.
- Zhang, H., Zhang, J., Lang, Z., Botella, J. R., & Zhu, J.-K. (2017). Genome editing—principles and applications for functional genomics research and crop improvement. *Critical Reviews in Plant Sciences, 36*(4), 291–309.
- Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics, 50*(2), 278–284.
- Zheng, J., Wu, H., Zhu, H., Huang, C., Liu, C., Chang, Y., et al. (2019). Determining factors, regulation system, and domestication of anthocyanin biosynthesis in rice leaves. *New Phytologist, 223*(2), 705–721.
- Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., et al. (2019). The population genetics of structural variants in grapevine domestication. *Nature plants, 5*(9), 965–979.
- Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nature Genetics, 51*(5), 865–876.
- Zhuang, W., Wang, X., Paterson, A. H., Chen, H., Yang, M., Zhang, C., et al. (2020). Reply to: Evaluating two different models of peanut's origin. *Nature Genetics, 52*(6), 560–563.

In-silico identification of small RNAs: a tiny silent tool against agriculture pest

Habeeb Shaik Mohideen¹, Kevina Sonawala¹ and Sewali Ghosh²

¹Bioinformatics and Entomoinformatics Lab, Department of Genetic Engineering, School of Bioengineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India, ²Department of Zoology and Advanced Biotechnology, Guru Nanak College, Chennai, Tamil Nadu, India

14.1 Introduction

Present-day agriculture faces losses from 20% to 40% of annual global crop production due to pest infestations (International Year of Plant Health, 2020). Many economically important crops are damaged by insects belonging to Coleoptera, Diptera, Hemiptera, Orthoptera, and Lepidoptera, adversely influencing crop production and yield. Traditional pest control practices mainly depend on extensive use of pesticides, which has led to pesticide resistance and the risk of toxic effects on nontarget organisms. Alternative practices like biopesticides, integrated pest management are comparatively ineffective in controlling some pests. On the other hand, biotechnological strategies like the expression of insecticidal Cry toxins in plants and RNAi technology have shown promising pest control results. The determination of small RNAs forms a promising approach towards pest control in many crops. This chapter aims to discuss small noncoding RNA (sncRNAs) and how their detection paves a way towards controlling agricultural pest infestations.

Any form of living organism responsible for causing a threat or damage to crops and livestock is considered an agricultural pest. Common agricultural pests include pathogens, nematodes, weeds, rodents, insects, and mites responsible for reducing agricultural productivity. Agronomically essential crops like wheat, rice, maize, soybean, potato, lentil, etc., are damaged and fed upon by insects belonging to Coleoptera, Diptera, Hemiptera, Orthoptera, and Lepidoptera order. Pests can attack anytime during or after the production cycle, thus hindering the crop's overall quality and yield (Savary et al., 2000).

During the late 1840s, one million people died due to the *Phytophthora infestans* fungus that caused the Irish potato famine. Over the years, crops like plum trees, grapevines, and olives were attacked by the bacterium *Xylella fastidiosa*, leading to about \$104 million losses in California and 180k of land Italy. The 2017 State of the World's Plants report ranked the top pests in the past five years based on The Centre for Agriculture and Biosciences International data. Cotton bollworm ranked first affecting plants like cotton and chickpea, followed by Tobacco whitefly, mainly affecting tomato and cotton crops. Two-spotted spider mite stood third, affecting tomato and common bean crops (State of the World's Plants, 2017). The 2019–20 Locust plague approximately infested 23 countries and damaged around 2.25 million hectares of land by April 2020.

Insect pests cause both direct and indirect types of injury to crops. When an insect's feeding and tunneling activities lead to harm or loss of plant parts, the resultant damage is considered direct crop damage. While in case of indirect damage, insects cause less harm but render the entry of various pathogens leading to different infections (Campbell & Reece, 2002). Thus directly or indirectly, pests have spawned enormous losses in the agriculture industry.

14.2 Small RNAs

Almost three decades ago, small RNA and its target gene were discovered in *Caenorhabditis elegans* (Lee, Feinbaum, & Ambros, 1993; Wightman, Ha, & Ruvkun, 1993). With this discovery soon began a new chapter of the small RNA

world in the book of noncoding RNAs (ncRNAs). ncRNAs are molecules that are not translated into proteins. Depending on their functions, they are divided into two major classes known as housekeeping RNAs and regulatory RNAs. Housekeeping RNAs comprise ribosomal RNA (rRNA), transfer RNA (tRNA), while regulatory RNAs include small ncRNAs (sncRNAs) and long ncRNAs (lncRNAs). These noncoding RNAs amount to the majority of the total RNA.

Tiny noncoding RNAs comprising approximately 18 to <200 nucleotides in length responsible for regulating gene expressions are known as sncRNAs. In recent years, various sncRNAs have been identified and characterized based on their biogenesis, interaction with different Argonaute family proteins, and lengths (Carthew & Sontheimer, 2009; Kim, Han, & Siomi, 2009). Some of these types of sncRNAs include piRNAs, miRNAs, siRNAs, and tRNA fragments.

14.3 Types of small noncoding RNAs

Piwi-interacting RNAs (piRNAs) are RNA molecules that interact with PIWI proteins belonging to the Argonaute family. These are approximately between 24–30 nucleotides in length and are generally grouped into 20–90 kilobases and form the most abundant class of sncRNA (Girard, Sachidanandam, Hannon, & Carmell, 2006). Apart from silencing transposable elements (TE), piRNAs are also involved in carrying out genomic and epigenetic regulations.

MicroRNAs (miRNAs) are approximately 22 nucleotides long, are responsible for posttranscriptional gene regulation via the Argonaute protein-aided repression and mRNA degradation. A single miRNA molecule can easily bind to multiple targets and vice versa, thus enabling diverse signaling patterns. Consequently, they are the most extensively researched sncRNA.

Small interfering RNA (siRNAs) are double-stranded RNA molecules responsible for degrading mRNAs to prevent translation and thus regulate the corresponding gene expression (Laganà et al., 2015). These are 20–27 bp molecules that play a vital role in cellular defense mechanisms against foreign genetic materials and TE via RNA interference mechanism.

14.4 Next-generation sequencing in agronomic advancements

The establishment of omics-based techniques has significantly influenced the enforcement of computational data mining tools leading to the gathering, integrating, and analyzing bioinformatics-based data. The sequencing of sizeable genomic information and yielding high throughput data is known as Next-generation sequencing (NGS). NGS technologies have provided solutions to various agricultural problems by allowing the genomic analysis of crops and livestock and understanding the complexity of various genetic interactions. NGS-based agrigenomics practices have contributed positively to the health, the yield of crops, and livestock, leading to increased productivity in the food, clothing, and pharmaceutical industries (Van Borm et al., 2014).

These approaches have also aided in understanding the genetic basis of agriculturally crucial traits and alteration of genes linked to the target phenotypic traits. Genomic selection, genome-wide association studies, and marker-assisted variants selection approaches have enhanced the breeding efficiency leading to increased production of nutritionally rich crops. With the help of NGS-aided technologies, the initiation, interaction, and elimination of plant diseases and disease etiology can be analyzed. Nowadays, deep sequencing technologies are utilized to predict and identify sncRNA and their targets in plants and pests to study corresponding defense mechanisms and develop appropriate strategies to control crop pests (Djami-Tchatchou, Sanan-Mishra, Ntushelo, & Dubery, 2017). For instance, in miRNAs, a range of bioinformatics-based algorithms and databases are preferred like miRBase, PicTar, RNAHybrid, TargetScan, miRanda, DIANA-microT-CDS, among others.

14.5 Small RNA world and their identification

14.5.1 MicroRNA

MicroRNA biogenesis is carried out via canonical and noncanonical pathways. Canonical biogenesis is initiated with the transcription of the primary-miRNA (pri-miRNA) transcript by RNA Polymerase II (Lucas & Raikhel, 2013). The microprocessor complex cleaves the pri-miRNA to produce the premature miRNA (pre-miRNA), which is exported to the cytosol by Exportin5 and processed to produce the mature miRNA duplex (Okada et al., 2009). Further, either strand of the mature miRNA duplex is incorporated into the proteins of the Argonaute (AGO) family to make up a miRNA-induced silencing complex (miRISC) (Yoda et al., 2010). The noncanonical pathways are divided into Dicer-

independent pathways and Droscha/PASHA-independent pathways. In the former, the shRNAs are cleaved by the microprocessor complex and exported to the cytoplasm via Exportin5, where they are again processed via AGO2-dependent cleavage (Yang et al., 2010). While in the latter case, m⁷G-pre-miRNA and mirtrons undergo cytoplasmic maturation mediated by Dicer endonuclease. M⁷G-pre-miRNA and mirtrons are exported by Exportin1 and Exportin5, respectively (Babiarz, Ruby, Wang, Bartel, & Blelloch, 2008; Ruby, Jan, & Bartel, 2007; Xie et al., 2013). Eventually, both the pathways lead to the formation of the miRISC complex. The degree of sequence complementarity among miRNAs and their targets, gene regulation is mediated by the miRNAs via mRNA decay, mRNA degradation, or translation inhibition (Fig. 14.1).

In 1993, *lin-4*, the first miRNA, and its target gene were discovered. It was responsible for negatively regulating the protein-coding gene *LIN-14* mRNA. Soon within a decade, various platforms relating to deep sequencing and identification started emerging. In 2002, miRBase was developed, and it soon became a crucial portal for the miRNA sequence repository for all species (Kozomara & Griffiths-Jones, 2014). Till now several algorithms have been developed to predict and identify miRNAs like MIRscan (Lim et al., 2003), triplet-SVM (Xue et al., 2005), MiPred (Jiang et al., 2007), miRDeep, miRCat, HHMMiR (Kadri, Hinman, & Benos, 2009), MatureBayes (Gkirtzou, Tsamardinos, Tsakalides, & Poirazi, 2010), miRNAFold (Tav, Tempel, Poligny, & Tahy, 2016), miReader (Jha & Shankar, 2013), miRPLEX (Mapleson, Moxon, Dalmay, & Moulton, 2013), miRIdentify (Hansen, Venø, Kjems, & Damgaard, 2014), deepSOM (Stegmayer, Yones, Kamenetzky, & Milone, 2017), and many more (Table 14.1).

MiRAlign works on *ab initio* algorithms to predict and detect miRNAs by aligning known pre-miRNAs' secondary structures (Wang et al., 2005). This tool aids in predicting species-specific as well as evolutionarily conserved miRNAs (Sewer et al., 2005). MiRseeker works by scanning euchromatic *Drosophila* sequences for conserved stem-loops and exhibiting the nucleotide divergence patterns of known miRNAs. It could detect most known *Drosophila* miRNAs and 48 new miRNAs, which were highly conserved in distant insect, nematode, or vertebrate genomes. Further, Lai et al. verified the expression of high-scoring 24 novel miRNA candidates by northern blotting (Lai, Tomancak, Williams, & Rubin, 2003). MiRanalyzer detects and predicts known and novel microRNAs by implementing an incredibly accurate machine learning algorithm. It can recall values of unseen data by 75% and hit 97.9% of the area under the curve values (Hackenberg, Sturm, Langenberger, Falcón-Pérez, & Aransay, 2009). The new substitute of this program, sRNAbench, is responsible for predicting small RNAs and their expression profiling, genome mapping, and studying other statistics. Analysis of sncRNA variants can also be carried out by sRNAbench (Rueda et al., 2015).

MiRDeep2 is a software package that identifies miRNAs with an accuracy of 98.6%–99.9%. It scrutinizes data from seven major animal clades with around 70% to 90% sensitivity. It can profoundly separate miRNAs from other argonaute-bound small RNAs, thus aiding in the accurate miRNA identification from Nematodes, common soil pests affecting crops (Friedländer, MacKowiak, Li, Chen, & Rajewsky, 2012). The UEA sRNA Toolkit's successor (Moxon et al., 2008), the UEA sRNA workbench (Stocks et al., 2012), comprises different Java-based tools for processing and analyzing small RNA NGS data. The MiRCat tool uses an sRNA dataset to predict the pre-miRNAs and mature miRNAs; it does so by detecting precursor miRNA hairpins. Tools4miRs is a one-stop solution for all the methods related to miRNA analysis. At present, it consists of 40 miRNA identification and 60 target prediction tools (Lukasik, Wójcikowski, & Zielonkiewicz, 2016). Various miRNA target genes can be predicted by programs like miRanda (John et al., 2004), RNA22 (Miranda et al., 2006), PITA (Kertesz, Iovino, Unnerstall, Gaul, & Segal, 2007), CleaveLand (Addo-Quaye, Miller, & Axtell, 2009), TargetsCan (Agarwal, Bell, Nam, & Bartel, 2015), miRDB (Wong & Wang, 2015) among others.

Spodoptera frugiperda is an annoying pest accountable for harming around 350 plant species causing extensive economic losses. Kakumani et al., detected 226 miRNAs in fall armyworm cell line Sf21; 116 candidates from these were found to be highly conserved in other pests like *Bombyx mori*, *Drosophila melanogaster*, and *Tribolium castenum*. They identified 110 miRNAs along with five miRNA clusters. Based on the computational analysis, miRNAs from *S. frugiperda* expressed higher homology than *B. mori* compared to other insects and pests like *D. melanogaster* and *T. castenum* (Kakumani et al., 2015). Rao et al., detected 58 miRNAs from *Spodoptera litura* based on secondary structure and sequence conservation analysis using *in silico* methods. They were further validated by experimental analysis; of these, 11 miRNAs manifested crucial changes in the developmental stages of the insects against which 128 possible target genes were predicted (Rao et al., 2012).

Meloidogyne incognita is a root-knot nematode that infects crop roots and drains their nutrients by parasitic behavior. Pests belonging to the *Meloidogyne* genus are accountable for 5% of global crop loss by affecting around 2000 plant types globally (Barker, Cathy Cameron Carter, & Sasser, 1985). Wang et al., with a computational pipeline approach, detected 44 unique miRNAs and seven miRNA clusters in *M. incognita*. MiR-100/let-7, miR-71-1/miR-2a-1, miR-71-2/miR-2a-2, and miR-279/miR-2b clusters are found to be conserved in other species, namely *C. elegans*, *A. suum*, *B. Malayi*, and *P. pacificus* (Wang et al., 2015).

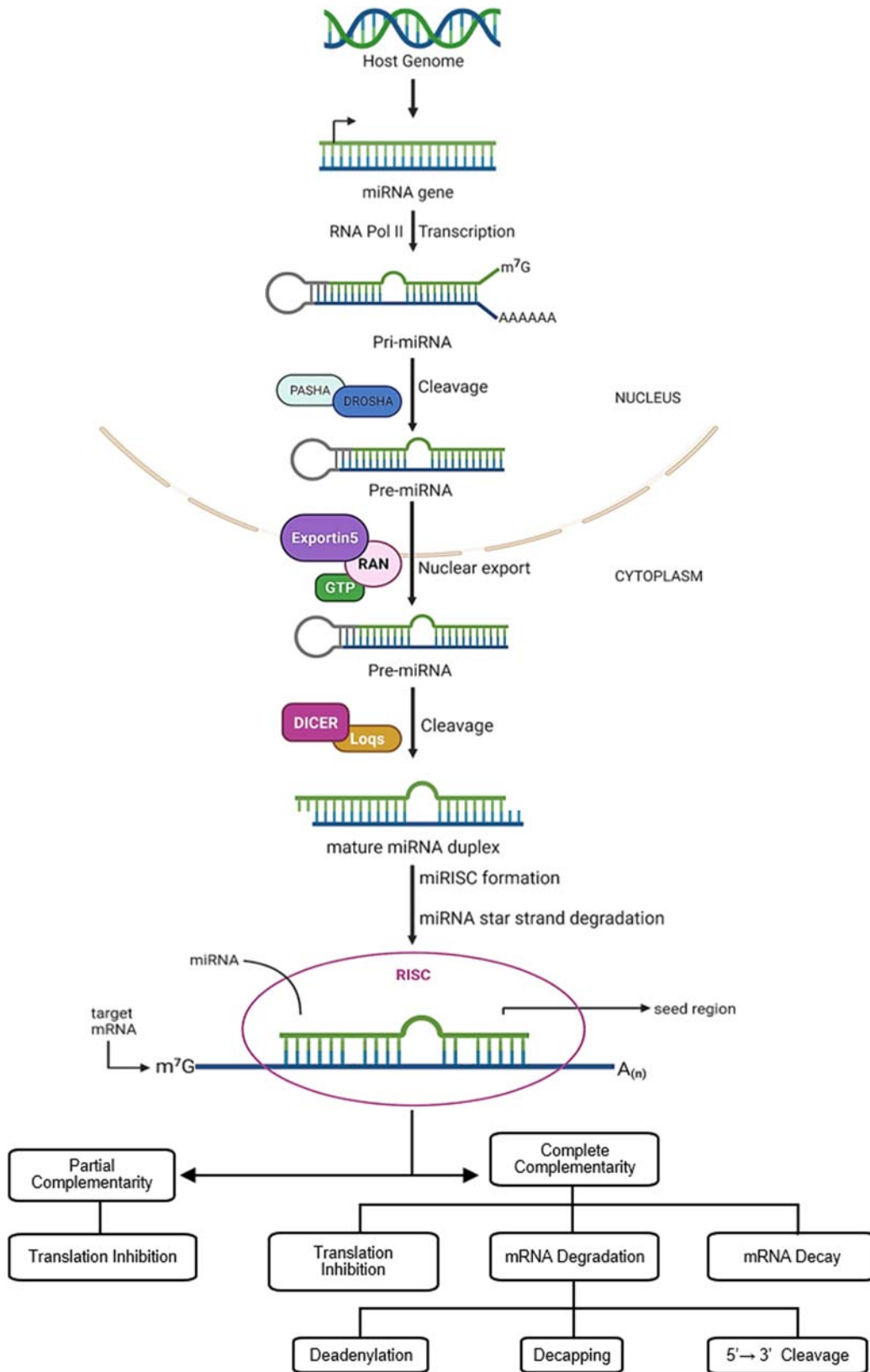


FIGURE 14.1 miRNA mediated mRNA regulation.

TABLE 14.1 Commonly used sRNA prediction tools.

Tool name	Salient feature	Tool link
Mirnov	A rapid, economical, highly parallelized, and multithreaded miRNA identification method.	http://wwwdev.ebi.ac.uk/enright-dev/mirnov/
MIRPIPE	A web platform serving miRNA homology detection and quantification.	https://github.molgen.mpg.de/pages/loosolab/www/software/Mirpipe/
sRNAtoolbox	A collection of sRNAbench and various sRNA downstream analysis tools.	https://bioinfo2.ugr.es/srnatoolbox/
miRIdentify	An open-source tool conducting read mapping and rigorous miRNA predictivity.	https://www.ncrnalab.dk/#mirdentify/mirdentify.php
mirTools 2.0	A tool aiding ncRNA detection, complete profiling, and functional annotation.	https://tools4mirs.org/software/all_in_one/mirtools-20/
miRDeep2	A highly sensitive sequence-based identification tool is covering seven major animal clades.	https://www.mdc-berlin.de/content/mirdeep2-documentation
UEA sRNA Workbench	A software package for MiRNA and SiRNA prediction, analysis, and thorough profiling.	http://srna-workbench.cmp.uea.ac.uk/mircat2/
MiPred	A random forest-based machine-learning algorithm that differentiates real pre-miRNA from false candidates.	http://server.malab.cn/MiPred/
mirDNN	A rapid and scalable deep ResNet with automated learning and high precision processing.	http://sinc.unl.edu.ar/web-demo/mirdnn/
deepSOM	An unsupervised deep model that accords a significant graphical navigation and result interpretation.	http://fich.unl.edu.ar/sinc/blog/web-demo/deepsom/
Prost!	A user-accessible tool that aids miRNA analysis of any given species.	https://prost.readthedocs.io/en/latest/

Menor et al., proposed a small RNA prediction approach based on the read’s nucleotide composition; thus, circumventing the need for reference genome or genomic information of other related species. As a result, a relatively a greater number of miRNAs can be detected using miRplex tool. Moreover, compared to piRNApredictor, with the help of this approach, it is possible to improve the true positive rate for piRNA by 60% (Menor, Baek, & Poisson, 2015).

InsectBase intends to provide a comprehensive platform for researchers who have an interest in analyzing insect genes. The database contains more than 12 million sequences, encompassing the genomes of 138 insects, transcriptomes of 116 insects, gene sets of 61 insects, 36 gene families of 60 insects, 7,544 miRNAs of 69 insects, 96,925 piRNAs from two insects (Yin et al., 2016).

14.5.2 PIWI-interacting RNAs

In a *Drosophila* germline, piRNAs were first reported as those sncRNAs transcribed from genetic elements like the Stellate locus and transposons (Aravin et al., 2001, 2003). The piRNA biogenesis is initiated by transcribing long ssRNAs by piRNA clusters, further fragmented by Zucchini/PLD6 (Ding et al., 2017) mediated cleavage to produce pre-piRNAs, which are then exported from the nucleus to the cytoplasm, where they undergo primary processing after binding to a PIWI protein. The 3’ ends of pre-piRNAs are shortened (Tang, Tu, Lee, Weng, & Mello, 2016) to a length characteristic of the receiving PIWI protein (Izumi et al., 2016; Kawaoka, Izumi, Katsuma, & Tomari, 2011). After the methylation of their 3’ ends, mature piRNAs enter the nucleus following their cleavage by PIWI proteins to induce transcriptional gene silencing (Brennecke et al., 2007). In the Ping pong pathway, secondary processing occurs when piRNA-directed slicing of target transcripts occurs by PIWI interactions, creating RNA fragments with 5’ monophosphate pre-pre-piRNA to PIWI protein and generate a secondary piRNA with ten nucleotides (Wang, Yoshikawa et al., 2014) complementary to the produced piRNA.

For the past two decades, just like miRNA prediction tools, several piRNA detections and target prediction tools were designed like piRpred, piRscan (Wu et al., 2018), 2L-piRNA (Liu, Yang, & Chou, 2017), pirnaPre (Yuan et al., 2016),

and so on. Common approaches to detect piRNA are based on immunoprecipitation techniques and deep sequencing in model organisms (Yin and Lin, 2007). Nevertheless, these approaches may have certain drawbacks, like piRNAs with low expression or those in which the ping pong model doesn't produce could go undetected (Das et al., 2008). Thus, in silico approaches that consider piRNAs' existing data to train in piRNA detection can provide a promising alternative approach in piRNA identification.

K-mers are certain k -tuples/ k -grams of nucleic acid sequences that can identify distinct regions within various biomolecules. The piRNAPredictor functions by predicting piRNAs based on a k -mer search algorithm. It's over 60% sensitive and has a precision rate above 90%. This type of method does not require a reference genome. Zhang et al., detected about 87536 piRNAs from the locust (Magor, Lecoq, & Hunter, 2008; Zhang, Wang, & Kang, 2011).

Wang et al., presented an algorithm that predicts piRNAs by analyzing their interactions with mobile elements. Its specificity, sensitivity, and accuracy are over 90%. *Chilo suppressalis* is a harmful rice pest that accounts for high yield loss (Seshu Reddy and Walker, 1990; Wang et al., 2020). With the help of Piano, about 82,639 novel piRNAs were predicted. The transposon targets of various species were also detected, including Asiatic rice borer, in which 44% piRNAs target SINE transposons and 42.4% target LINE transposons (Wang, Liang et al., 2014).

Rosenkranz et al., and Jung et al., presented deep sequencing-based approaches to detect piRNA clusters (Jung, Park, & Kim, 2014; Rosenkranz and Zischler, 2012). Finally, Brayet et al., proposed piRPred, which detects piRNAs depending on the telomere/centromere vicinity apart from other general features of the piRNA algorithm. The algorithm's machine learning method is established on multiple kernels and a support vector machine (SVM) classifier (Brayet, Zehraoui, Jeanson-Leh, Israeli, & Tahi, 2014).

14.5.3 Small interfering RNAs

In 1999, siRNAs were first reported by Hamilton et al., as part of posttranscriptional gene silencing in plants (Hamilton and Baulcombe, 1999). siRNA biogenesis is initiated with the formation of dsRNA via transcription by RNA polymerase II or III. RBD domain of Dicer complex recognizes and cleaves the dsRNA into short fragments with two nucleotides 3' overhang (siRNAs duplexes). They further bind to AGO protein and separate into single strands, the protein-bound strand is called the guide strand, and the other is called the passenger strand, which is ejected. This protein RNA complex is integrated into an active RISC by the activity of the RISC-Loading complex. To identify a stable terminus of the siRNA, R2D2 carries tandem dsRNA binding domains, while Dicer-2 deals with the other less stable extremity. The MID domain of AGO recognizes the stable end of the duplex. Thus, the sense strand whose 5' end is discarded by MID is ejected. After forming mature RISC, siRNAs base-pair to their target mRNA and cleaves it to prevent it from being used as a translation template (Bartel, 2005; Kim et al., 2009; Xia, Mao, Paulson, & Davidson, 2002).

The siRNAs can be predicted and identified from tools like siRNA-Finder, MysiRNA, NATpare, and some tools, as mentioned in Table 14.1. A relatively lesser number of identification algorithms have been developed for siRNA as compared to miRNA and piRNA. More light is thrown on tools relating to target-specific siRNA designing, efficacies, and prediction of targets and off-targets. SiRNadb is a siRNA database designed to aid researchers in determining which siRNA can be used to inhibit their gene of interest. It also gives information on siRNAs with known efficacy and the ones predicted to exhibit high efficacy, thermodynamic properties of siRNAs, the potential for sequence-related off-target effects (Chalk, Warfinge, Georgii-Hemming, & Sonhammer, 2005).

Choudhary et al., identified 16 favorable siRNAs in six *Helicoverpa armigera* hormonal pathway genes out of over 2000 detected siRNAs. Old World bollworm is a significant pest that affects vital crops like cotton, chickpea, tomato, sorghum, etc. Therefore these siRNAs are potential candidates targeting hormone biosynthesis and eventually disrupting the insect life cycle (Choudhary and Sahi, 2011).

14.6 Limitations

Some possible drawbacks can be associated with the prediction, detection, and validation of sncRNAs. For instance, problems associated with the transcript orientation, determination, and involvement of processing sites within hairpin sequences. Prediction of false-positive sncRNA candidates. Lack of experimental validation due to various factors like a failed expression of sncRNA associated genes, the unknown expression pattern of the predicted RNA/targets, etc. Instances like these may cause hindrances in the computational identification and validation processes of sncRNAs (Aravin and Tuschl, 2005).

14.7 Conclusion

With a surging global population, the need for food and textile has increased like never before. It has burdened the agricultural industry with producing higher yields of crops and livestock products annually, which has soared the production and use of pesticides and herbicides to newer levels to combat extensive damages caused by crop pests and weeds. As of November 2019, about 2 million tonnes of pesticides are exploited annually on a global scale (Sharma et al., 2019). The conventional practices to control pests have played significant roles in imparting pesticide resistance to various agricultural pests. However, such immense use of pesticides on large scales has adversely affected the environment causing toxic effects to nontargeted flora and fauna. It is also leading to soil and water pollution, thus terribly affecting the respective ecosystems. This calls for alternate measures to control crop pests in a way that imparts minimum damage to the environment.

Among the upcoming technologies, RNA-based pest control strategies have gained potential. RNAi was first reported in *Petunia* and described in *C. elegans* (Fire et al., 1998). Effective RNAi responses are seen in pests in the orders of Orthoptera, Coleoptera, and while pests belonging to Lepidoptera, Diptera, and Hemiptera depict relatively

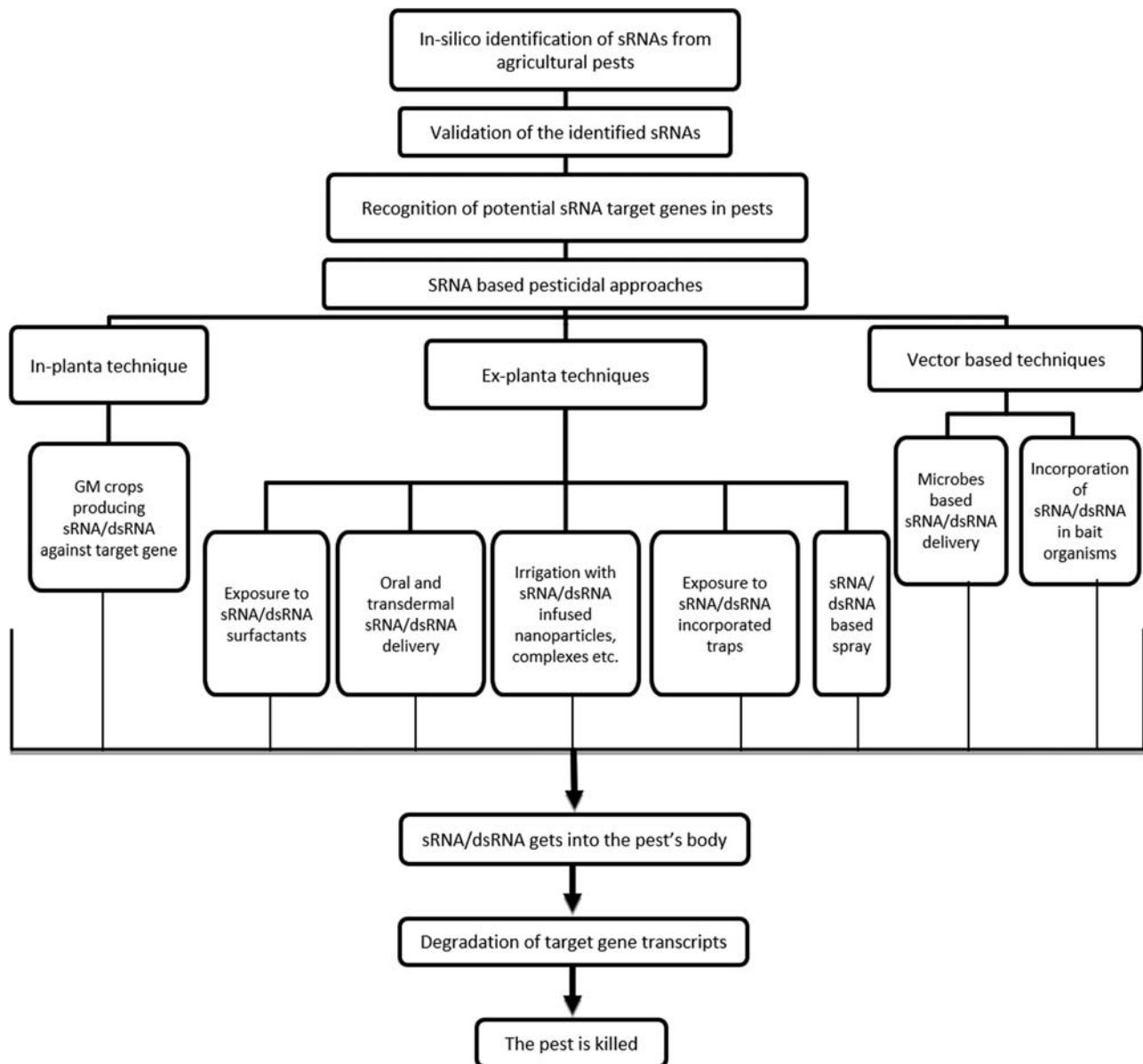


FIGURE 14.2 Applications of in silico sRNA identification tools for pest control.

lower responses (Cooper, Silver, Zhang, Park, & Zhu, 2019; Wynant, Santos, & Vanden Broeck, 2014; Xu et al., 2016). Small RNA-based pesticidal approaches involve *in-planta* techniques like the expression of dsRNA in transgenic crops and the use of plant-incorporated protectants. The *ex planta* approach involves the direct application of dsRNA as an insecticide, for instance in the form of sprays.

To target any crop-specific pests, it is crucial to have information on the target genes and their role in the insects' physiology. Furthermore, the corresponding small RNA can be incorporated into the pest via RNA-based approaches. Thus identifying small RNAs and their targets plays a fundamental role in any RNA-based pesticidal procedures. Moreover, NGS techniques and modern computational tools give an upper hand in the prediction and identification processes of sncRNA and their targets which would have been tedious and time-consuming in conventional approaches. (Fig. 14.2)

Acknowledgments

We acknowledge and thank DST-SERB, India for the grant under the young scientist scheme: file#- YSS/2014/000293. Authors thank the administration and management of SRM IST for all the support provided.

References

- Addo-Quaye, C., Miller, W., & Axtell, M. J. (2009). *Bioinformatics (Oxford, England)*, 25, 130.
- Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). *Elife*, 4.
- Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., . . . Tuschl, T. (2003). *Developmental Cell*, 5, 337.
- Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M., & Gvozdev, V. A. (2001). *Current Biology: CB*, 11, 1017.
- Aravin, A., & Tuschl, T. (2005). *FEBS Letters*, 579, 5830.
- Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P., & Blelloch, R. (2008). *Genes & Development*, 22, 2773.
- Barker, K. R., Cathy Cameron Carter., & Sasser, J. N. (1985). *An Advanced treatise on meloidogyne (Book, 1985) [WorldCat.org]*. Raleigh, N.C., U.S.A.: Dept. of Plant Pathology, North Carolina State University.
- Bartel, B. (2005). *MicroRNAs directing siRNA biogenesis* (Vol. 12, pp. 569–571). Nature Publishing Group.
- Brayet, J., Zehraoui, F., Jeanson-Leh, L., Israeli, D., & Tahri, F. (2014). *Bioinformatics*. Oxford University Press.
- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). *Cell*, 128, 1089.
- Campbell, N. A., & Reece, J. B. (2002). *Campbell & Reece, Biology | Pearson* (6th (ed.)). Pearson.
- Carthew, R. W., & Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell*, 642–655.
- Chalk, A. M., Warfinge, R. E., Georgii-Hemming, P., & Sonnhammer, E. L. L. (2005). *Nucleic Acids Research*, 33, D131.
- Choudhary, M., & Sahi, S. (2011). *Indian Journal of Experimental Biology*, 49, 469.
- Cooper, A. M. W., Silver, K., Zhang, J., Park, Y., & Zhu, K. Y. (2019). *Pest Management Science*, 75, 18.
- Das, P. P., Bagijn, M. P., Goldstein, L. D., Woolford, J. R., Lehrbach, N. J., Sapetschnig, A., . . . Miska, E. A. (2008). *Molecular Cell*, 31, 79.
- Ding, D., Liu, J., Dong, K., Midic, U., Hess, R. A., Xie, H., . . . Chen, C. (2017). *PNLDC1 is essential for piRNA 3' end trimming and transposon silencing during spermatogenesis in mice*, 8, 1–10.
- Djami-Tchatchou, A. T., Sanan-Mishra, N., Ntushelo, K., & Dubery, I. A. (2017). *Functional roles of microRNAs in agronomically important plants-potential as targets for crop improvement and protection* (Vol. 8, p. 378). Frontiers Research Foundation.
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). *Nature*, 391, 806.
- Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). *Nucleic Acids Research*, 40, 37.
- Girard, A., Sachidanandam, R., Hannon, G. J., & Carmell, M. A. (2006). *Nature*, 442, 199.
- Gkirtzou, K., Tsamardinos, I., Tsakalides, P., & Poirazi, P. (2010). *PLoS One*, 5, e11843.
- Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J. M., & Aransay, A. M. (2009). *Nucleic Acids Research*, 37, W68.
- Hamilton, A. J., & Baulcombe, D. C. (1999). *Science*, 286, 950.
- Hansen, T. B., Venø, M. T., Kjems, J., & Damgaard, C. K. (2014). *Nucleic Acids Research*, 42, 124.
- International Year of Plant Health 2020 FAO Food and Agriculture Organization of the United Nations.
- Izumi, N., Shoji, K., Sakaguchi, Y., Honda, S., Kirino, Y., Suzuki, T., . . . Tomari, Y. (2016). *Cell*, 164, 962.
- Jha, A., & Shankar, R. (2013). *PLoS One*, 8, e66857.
- Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). *Nucleic Acids Research*, 35.
- John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. (2004). *PLoS Biology*, 2, e363.
- Jung, I., Park, J. C., & Kim, S. (2014). *Computational Biology and Chemistry*, 50, 60.
- Kadri, S., Hinman, V., & Benos, P. V. (2009). *BMC Bioinformatics, BioMed Central*, S35.
- Kakumani, P. K., Chinnappan, M., Singh, A. K., Malhotra, P., Mukherjee, S. K., & Bhatnagar, R. K. (2015). *PLoS One*, 10.
- Kawaoka, S., Izumi, N., Katsuma, S., & Tomari, Y. (2011). *Molecular Cell*, 43, 1015.
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., & Segal, E. (2007). *Nature Genetics*, 39, 1278.
- Kim, V. N., Han, J., & Siomi, M. C. (2009). *Biogenesis of small RNAs in animals* (Vol. 10, pp. 126–139). Nature Publishing Group.

- Kozomara, A., & Griffiths-Jones, S. (2014). *Nucleic Acids Research*, 42, D68.
- Laganà, A., Veneziano, D., Russo, F., Pulvirenti, A., Giugno, R., Croce, C. M. A., & Ferro, A. (2015). *Methods Mol. Biol.*, 1269, 393.
- Lai, E. C., Tomancak, P., Williams, R. W., & Rubin, G. M. (2003). *Genome Biology*, 4, R42.
- Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). *Cell*, 75, 843.
- Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., . . . Bartel, D. P. (2003). *Genes & Development*, 17, 991.
- Liu, B., Yang, F., & Chou, K. C. (2017). *Mol. Ther. - Nucleic Acids*, 7, 267.
- Lucas, K., & Raikhel, A. S. (2013). *Insect MicroRNAs: Biogenesis, expression profiling and biological functions* (Vol. 43, pp. 24–38). Pergamon.
- Lukasik, A., Wójcikowski, M., & Zielenkiewicz, P. (2016). *Bioinformatics (Oxford, England)*, 32, 2722.
- Magor, J. I., Lecoq, M., & Hunter, D. M. (2008). *Crop Protection (Guildford, Surrey)*, 27, 1527.
- Mapleson, D., Moxon, S., Dalmay, T., & Moulton, V. (2013). *J. Exp. Zool. Part B Mol. Dev. Evol.*, 320, 47.
- Menor, M. S., Baek, K., & Poisson, G. (2015). *Int. J. Mol. Sci.*, 16, 1466.
- Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., . . . Rigoutsos, I. (2006). *Cell*, 126, 1203.
- Moxon, S., Schwach, F., Dalmay, T., MacLean, D., Studholme, D. J., & Moulton, V. (2008). *Bioinformatics (Oxford, England)*, 24, 2252.
- Okada, C., Yamashita, E., Lee, S. J., Shibata, S., Katahira, J., Nakagawa, A., . . . Tsukihara, T. (2009). *Science*, 326, 1275.
- Rao, Z., He, W., Liu, L., Zheng, S., Huang, L., & Feng, Q. (2012). *PLoS One*, 7, e37730.
- Rosenkranz, D., & Zischler, H. (2012). *BMC Bioinformatics*, 13, 5.
- Ruby, J. G., Jan, C. H., & Bartel, D. P. (2007). *Nature*, 448, 83.
- Rueda, A., Barturen, G., Lebrón, R., Gómez-Martín, C., Alganza, Á., Oliver, J. L., & Hackenberg, M. (2015). *Nucleic Acids Research*, 43, W467.
- Savary, S., Willocquet, L., Elazegui, F. A., Teng, P. S., Van Du, P., Zhu, D., . . . Srivastava, R. K. (2000). *Plant Disease*, 84, 341.
- Seshu Reddy, K. V., & Walker, P. T. (1990). *International Journal of Tropical Insect Science*, 11, 563.
- Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M. J., . . . Zavolan, M. (2005). *BMC Bioinformatics*, 6, 267.
- Sharma, A., Kumar, V., Shahzad, B., Tanveer, M., Sidhu, G. P. S., Handa, N., . . . Thukral, A. K. (2019). *SN Applied Sciences*, 1.
- State of the World's Plants 2017 Royal Botanic Gardens, Kew.
- Stegmayer, G., Yones, C., Kamenetzky, L., & Milone, D. H. (2017). *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14, 1316.
- Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., . . . Moulton, V. (2012). *Bioinformatics (Oxford, England)*, 28, 2059.
- Tang, W., Tu, S., Lee, H. C., Weng, Z., & Mello, C. C. (2016). *Cell*, 164, 974.
- Tav, C., Tempel, S., Poligny, L., & Tahi, F. (2016). *Nucleic Acids Research*, 44, W181.
- Van Borm, S., Belák, S., Freimanis, G., Fusaro, A., Granberg, F., Höper, D., . . . Rosseel, T. (2014). *Methods in Molecular Biology*, 1247, 415.
- Wang, K., Liang, C., Liu, J., Xiao, H., Huang, S., Xu, J., & Li, F. (2014). *BMC Bioinformatics*, 15, 1.
- Wang, W., Yoshikawa, M., Han, B. W., Izumi, N., Tomari, Y., Weng, Z., & Zamore, P. D. (2014). *Molecular Cell*, 56, 708.
- Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., & Li, Y. (2005). *Bioinformatics (Oxford, England)*, 21, 3610.
- Wang, Y., Huang, C., Hu, B., Liu, Y., Walter, G. H., & Hereward, J. P. (2020). *Pest Management Science*, 76, 695.
- Wang, Y., Mao, Z., Yan, J., Cheng, X., Liu, F., Xiao, L., . . . Xie, B. (2015). *PLoS One*, 10.
- Wightman, B., Ha, I., & Ruvkun, G. (1993). *Cell*, 75, 855.
- Wong, N., & Wang, X. (2015). *Nucleic Acids Research*, 43, D146.
- Wu, W. S., Huang, W. C., Brown, J. S., Zhang, D., Song, X., Chen, H., . . . Lee, H. C. (2018). *Nucleic Acids Research*, 46, W43.
- Wynant, N., Santos, D., & Vanden Broeck, J. (2014). *International Review of Cell and Molecular Biology* (pp. 139–167). Elsevier Inc.
- Xia, H., Mao, Q., Paulson, H. L., & Davidson, B. L. (2002). *Nature Biotechnology*, 20, 1006.
- Xie, M., Li, M., Vilborg, A., Lee, N., Di Shu, M., Yartseva, V., . . . Steitz, J. A. (2013). *Cell*, 155, 1568.
- Xu J., Wang X.-F., Chen P., Liu F.-T., Zheng S.-C., Ye H., Mo M.-H., (2016). mdpi.com.
- Xue, C., Li, F., He, T., Liu, G. P., Li, Y., & Zhang, X. (2005). *BMC Bioinformatics*, 6, 310.
- Yang, S., Maurin, T., Robine, N., Rasmussen, K. D., Jeffrey, K. L., Chandwani, R., . . . Lai, E. C. (2010). *Proceedings of the National Academy of Sciences of the United States of America*, 107, 15163.
- Yin, C., Shen, G., Guo, D., Wang, S., Ma, X., Xiao, H., . . . Li, F. (2016). *Nucleic Acids Research*, 44, D801.
- Yin, H., & Lin, H. (2007). *Nature*, 450, 304.
- Yoda, M., Kawamata, T., Paroo, Z., Ye, X., Iwasaki, S., Liu, Q., & Tomari, Y. (2010). *Nature Structural & Molecular Biology*, 17, 17.
- Yuan, J., Zhang, P., Cui, Y., Wang, J., Skogerbø, G., Huang, D.-W., . . . He, S. (2016). *Bioinformatics (Oxford, England)*, 32, 1170.
- Zhang, Y., Wang, X., & Kang, L. (2011). *Bioinformatics (Oxford, England)*, 27, 771.

This page intentionally left blank

Section II

Omics application

This page intentionally left blank

Bioinformatics-assisted multiomics approaches to improve the agronomic traits in cotton

Sidra Aslam¹, Muhammad Aamer Mehmood¹, Mehboob-ur Rahman^{2,3}, Fatima Noor¹ and Niaz Ahmad^{2,3}

¹Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Faisalabad, Pakistan, ²Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering, Faisalabad, Pakistan, ³Department of Biotechnology, Pakistan Institute of Engineering & Applied Sciences (PIEAS), Islamabad, Pakistan

15.1 Introduction

15.1.1 A bird's-eye view of the world cotton market

Economically, cotton is considered a valuable - cash crop worldwide. It is a leading income source for more than one billion people globally. According to an estimate, the annual cotton production in the world is about 25 Mtons, with an overall worth of nearly 12 billion dollars (Khan et al., 2020). Cotton is cultivated in around a hundred countries on an area of over 30 M hectares worldwide (Townsend, 2020; Zaib et al., 2020).

The maximum share in the total cotton production comes from Asian countries including China, India, Pakistan, etc., followed by America, Europe, and Africa (Jabran, UI-Allah, Chauhan, & Bakhsh, 2019). In the year 2017, major cotton exporters were India (7.6 billion dollars), China (15.1 billion dollars), and the US (7.6 billion dollars), while the major cotton importers were Bangladesh (5.3 billion dollars), Vietnam (4.2 dollars), and China (8.6 billion dollars) (Khan et al., 2020). China, United States, Pakistan, India, Turkey, Brazil, Burkina Faso, Australia, Uzbekistan, and Turkmenistan are considered as the top ten cotton producers (Khan et al., 2020). The world textile industry relies on cotton for the production of garments, with an estimated volume of 748 billion in 2016 (Voora, Larrea, & Bermudez, 2020). Due to its huge economic importance, it's also known as “white gold.”

15.1.2 An overview of omics mainly focused on plant-omics

The term “omic” is originated from the Latin suffix “ome” meaning mass or many. Omics refers to the unified technologies used to explore the roles, behavior, and relationship of the various types of molecules that make up the cells of an organism (Datta, 2017). Different omics tools help understanding the main differences in proteins, RNA, DNA, and many other cellular molecules present among different individuals of a species. Hence, it can be inferred that omics is an integrated field of genomics, transcriptomics, metabolomics, and proteomics. All these approaches ultimately lead to the identification of the total number of genes, metabolites, mRNA, and proteins in a holistic way. Likewise, the omics technology is also helpful in understanding disease etiology through the process of diagnosis, screening, prognosis, and biomarker discovery (Poisot, Péquin, & Gravel, 2013).

Plant-omics is a fast-growing, effective, and vital field of study. There are a lot of new omics technologies such as microarrays, RNA-Seq, proteomics, SNP genotyping, and NMR which are used to unravel the genetic circuits of various biological functions. The characterization, detection, identification, and cell metabolic profiles of living organisms under certain environmental circumstances, the fast-growing term metabolomics is used (Kumar, Kuzhiumparambil, Pernice, Jiang, & Ralph, 2016). Transcriptome has been considered an invaluable tool to understand and predict gene functions (Kobayashi, Ohyanagi, & Yano, 2014). Huge data on functional and structural changes within the cell can be

produced by omics experiments conducted by high throughput assays. These advanced methods facilitated the perception of molecular responses to tissue and cell damage that also helped in the understanding of functional cellular systems (Aardema & MacGregor, 2003).

Fitting a crop cultivar to a particular environment is the main challenge for plant researchers. Moreover, it exposes the reality that genotype only will not be enough to support the biotechnology-driven crop improvement program, but a combination of omics technologies are required to provide reliable information (Sirangelo, 2019). The analysis of large datasets of multiomics can only be performed computationally. Many packages and software are available for the better processing of voluminous omics data, but the visualization of these large datasets remains a crucial task for bioinformaticians (Shaheen, Iqbal, & Zafar, 2016).

15.1.3 Introduction of bioinformatics in the area of next-generation sequencing

All branches of biological sciences that depend on nucleic acid sequence data have been profoundly changed in the last few decades, driven by the advent of next-generation sequencing (NGS) technologies. The NGS technologies offer high-throughput methods to investigate the sequences of nucleic acids and have become a most important and valuable tool in the applications of the life sciences (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013).

Compared to conventional Sanger sequencing, the NGS allows millions of bases to be sequenced at once at a relatively low cost (Metzker, 2010). The impact of NGS is egalitarian in that it allows both small and large research groups to solve problems in the field of biology and genetics including those in agriculture, virology, forensic science, and plant biology. Moreover, this technology is developing in parallel with the online availability of a variety of biological data, which makes it possible to address a kind of question never possible before. Bioinformatics approaches and web databases are needed for effective use of genetic, proteomic, transcriptomic, and metabolomic data important in enhancing the crop yield. With the improvement in technology and biological data, we are quickly moving to that point where not only the “model plants” but every plant is “open” to the power of NGS technology applications (Egan, Schlueter, & Spooner, 2012).

Bioinformatics is providing multiple tools for analyzing the NGS data, ranging from short-read alignment programs to algorithms for the recognition of structural variations (Sripathi et al., 2016). Concerning the growing challenges of NGS data storage, analysis, and interpretation, bioinformatics is increasingly becoming the rate-limiting step for NGS inclusion into translational research (Pereira, Oliveira, & Sousa, 2020). While using the NGS platforms, there is a minimum of four levels of genomic sequence analysis to consider (Horner et al., 2009; Schlötterer, 2002). The first step is the generation of DNA/RNA sequence reads. For this, sequencing devices integrated with software are used to convert the raw sequencing signals into bases of short nucleotide reads associated with the base quality score. Research laboratories face the problem of computer resources in the storage of raw signal and sequencing files as short read collection in FASTQ format. Safe storage of raw sequences is required for bioinformatics analysis. NGS technologies produced raw DNA data that can be submitted to the sequence read archive database of NCBI, while mRNA-Seq can be submitted to the Gene expression omnibus database (Hong et al., 2013).

The second step consists of contigs and scaffold alignment and assembly, which help to detect variants. The requirements for sequence alignment and variant detection depend on the NGS project format complexity (El-Metwally, Hamza, Zakaria, & Helmy, 2013). Short reads from small genomes are less complex than the large genomes of higher plants which make it easier to compute, align and assemble. Transfer of preedited sequencing data in the proper format to a software of alignment, assembly, and variant detection is usually straightforward. Moreover, many free software and packages offer to perform such kinds of tasks (Kulski, 2016).

The third step is the integration and visualization of assembled sequences. A lot of bioinformatics platforms are available to virtually transcribe, translate as well as annotate the nucleotide sequences to an advanced informatics level, like defining coding and noncoding regions, untranslated regions, repeat elements, and signal peptides. Five categories of annotation software have been described by Yandell and Ence (Yandell & Ence, 2012): (1) *ab initio* and evidence-based gene predictors; (2) aligners and assemblers for protein and RNA-Seq; (3) pipelines of genome annotation; (4) selectors and combiners; (5) genome browsers for data curation. NCBI provides a typical pipeline of genome annotation, while BUSCO, babelomics, PASA, MEGANTE, and MAKER are also considered important genome annotation tools (Kulski, 2016). The fourth step includes the consolidation of data, from different platforms of NGS, into one single bioinformatics output with accessible tools and web addresses. Integration and visualization of annotated data are done with genome browsers like those displayed at JBrowse, Ensemble UCSC, and genome maps (Medina et al., 2013).

15.1.4 Brief description of “integration of omics”

Plant system biology focuses on the understanding of interrelationships among plant genotypes and the corresponding phenotypes which allow the interpretation of the proteins or genes all at once (Greenbaum, Luscombe, Jansen, Qian, & Gerstein, 2001). Investigation of the complicated biological processes is not only important to understand the gene's function but also to determine the interrelationship between different metabolic pathways (Schaal, 2019).

Advancement in the high throughput techniques based on nuclear magnetic resonance (NMR) and mass spectrometry has allowed the detection of a broad spectrum of small molecules. Metabolite profiling depends on the investigation of the largest group of metabolites involved in particular metabolic pathways (Wolfender, Marti, Thomas, & Bertrand, 2015). It has been investigated the metabolites profiling had also been served as a method of diagnosing specific genotypes (Fernie, Trethewey, Krotzky, & Willmitzer, 2004). Furthermore, it also enables the detection of biotic and abiotic responses in a plant (Urano et al., 2009), as well as the determination of the function of unidentified genes (Saito, Hirai, & Yonekura-Sakakibara, 2008). For a complete study of plant metabolic pathways, detection of an enzymatic gene is not anymore acceptable, hence other omics fields include proteomics, genomics, and transcriptomics can also empower with additional details to rigorously decipher the plant metabolic pathways (Oksman-Caldentey & Saito, 2005; Yuan, Galbraith, Dai, Griffin, & Stewart, 2008).

Integration of omics has also evaluated the information flow obtained from one omics to others (Buescher & Driggers, 2016). Therefore the objective of multiomics data integration is to combine the different types of data to build a model that can be used to predict the composite traits and phenotypes (Fig. 15.1)

The main purpose behind the integration of omics is to remove the gap among the generation of data as well as the capability to study and explore the complex biological mechanism. Omics also makes it possible to identify the biomarkers and the relationships among datasets that have not been examined (Rajasundaram & Selbig, 2016). Another main benefit of the integration of omics is that there are fewer chances of false positives that are produced from the single-source dataset (Misra, Langefeld, Olivier, & Cox, 2019).

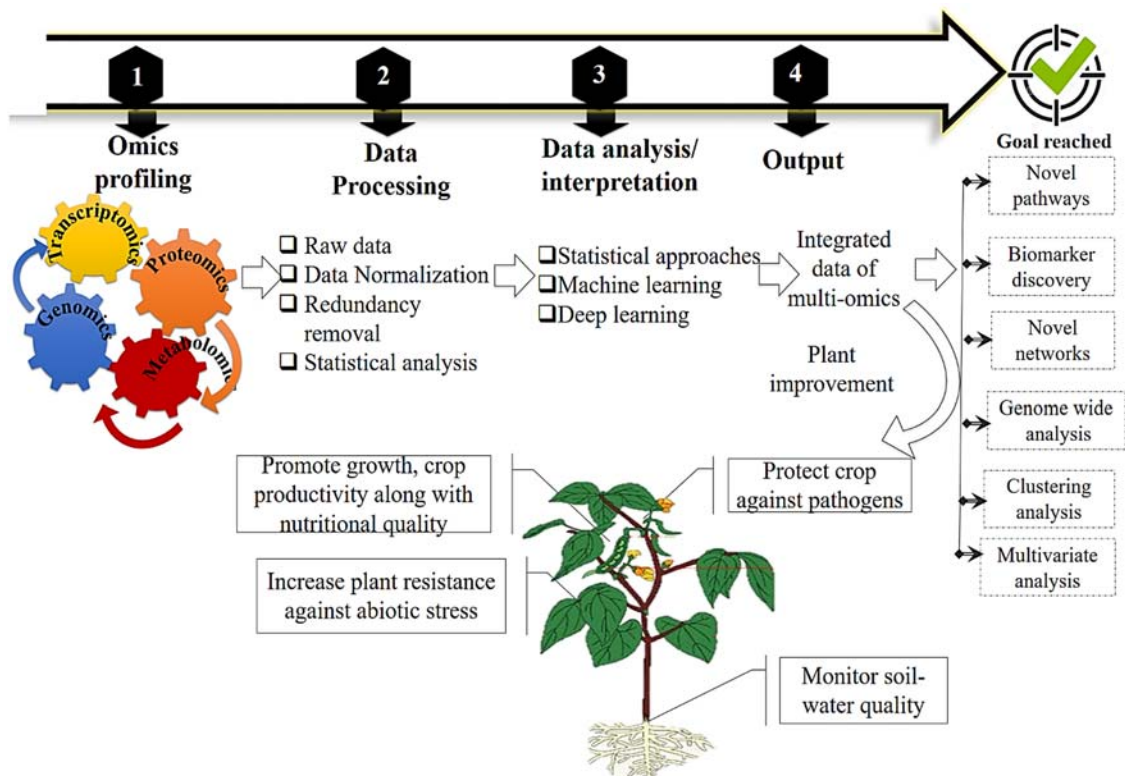


FIGURE 15.1 The workflow of multiomics data integration.

15.1.5 Why is multiomics study preferred over single-omics?

As compared to the studies of single-omics, multiomics techniques provide the opportunity to deeply understand the biological mechanisms involved in plant growth, and disease development (Pinu et al., 2019). The adoption of multiomics strategies has become a common practice in many fields of biology (Fig. 15.2). Therefore scientists are now focusing on the development of far-reaching multiple omics experimental methods and also trying the integration of different datasets to achieve a deep understanding of various biological functions. Appropriate integration of multiple omics data makes it easier to comprehensively examine the biological pathways. For example, it is possible to know how a given genotype affects its phenotype, as well as to characterize the molecular mediators that control the principal mechanisms. Multiomics also made it possible to find the important biological pieces of evidence in pathways that otherwise cannot be clarified with the single-omics methods alone (Hasin, Seldin, & Lulis, 2017).

Integrated approaches have also great importance in plant sciences and helped a lot in understanding the mechanisms involved in plant senescence, diseases, or plant responses to stress (Großkinsky, Syaifullah, & Roitsch, 2018). New practical methods and software to integrate the multiomics datasets are needed to explore and analyze the complex pathways in plants. Machine learning methods are also necessary for the model-based interpretation of multiple omics data for pathways analysis (Graw et al., 2021).

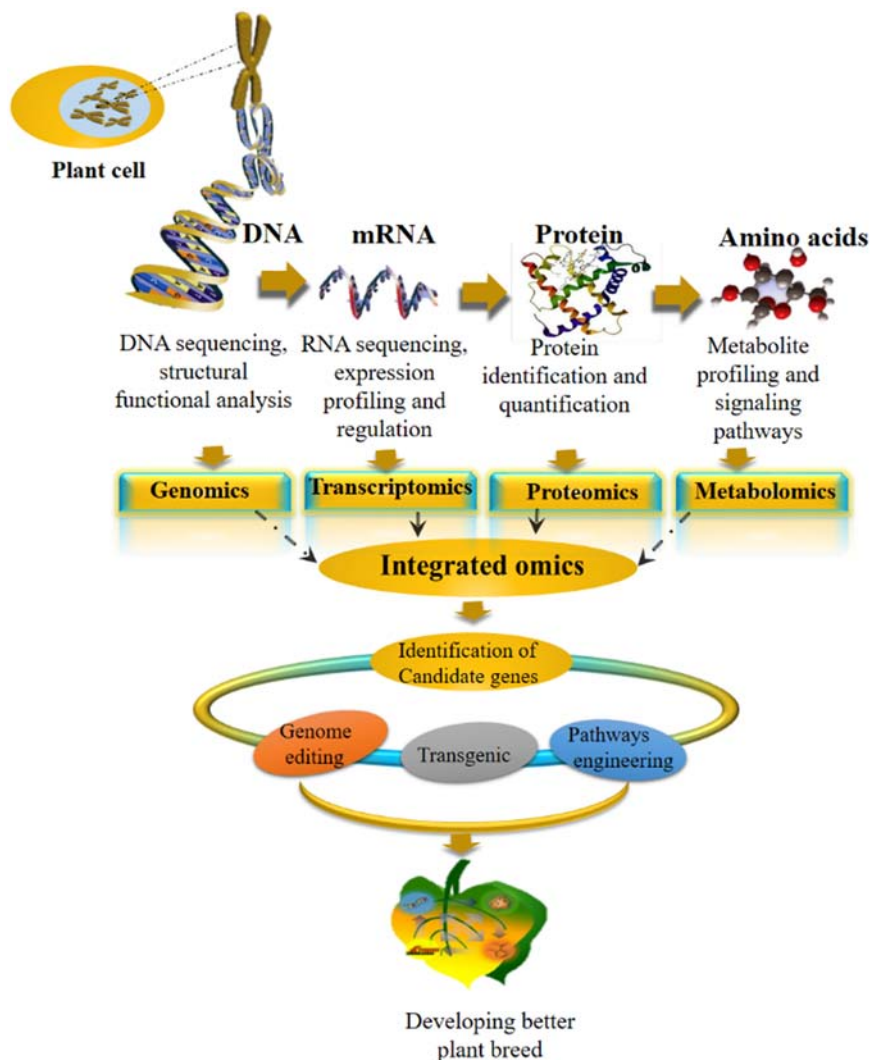


FIGURE 15.2 A schematic diagram represents the integration of omics approaches.

15.2 Big data in biology and omics

System biology revolves around transcriptomics, metabolomics, and proteomics which together offer extensive information regarding the expression level of the associated transcripts, metabolites, and proteins, respectively (Aizat, Goh, & Baharum, 2018). A huge amount of data produced from these platforms often have no connection among them. For example, it is not possible to compile several hundreds of millions of transcripts with their corresponding protein and metabolic pathways. Indeed, the backbone of omics research serves as a crucial human resource/biological component for the integration and processing of data (Palsson & Zengler, 2010). Considerably, multiomics techniques make it easier to view and understand the specific biological processes in terms of the whole plant as well as their environmental interaction (Brauer, Singh, & Popescu, 2014).

In the present era of omics, data is being presented in different ways, representing information in different domains including proteome, genome, epigenome, transcriptome, and metabolome. A plurality of distinct tools and approaches are available to handle or analyze the big data and allow translating the big data into meaningful knowledge which thereafter produces a potential bonanza. Big data biology is a field of data-intensive science so-called omics biology. This discipline was proposed based on the availability of a large amount of omics data. Big data in biology facilitates omics analysis to solve biology-related problems. Nowadays, big data is considered a burning issue in the scientific community, but the understanding might be potentially tricky or sometimes misleading. The word itself refers to the massive amount of data, representing a particular aspect (Afendi et al., 2013; Schatz, 2012). Plant Omics is a rapidly emerging area in the scientific community because it became a need of the hour to tackle the major concerns that the world has been confronted regarding agriculture. Various studies have been made to understand the physiological, ultra-structural, and molecular mechanisms of the plant responses in terms of abiotic stresses. Such kind of investigation has been performed which covered various omics approaches such as proteomics, genomics, metagenomics, metaproteomics, metatranscriptomics, and metabolomics that reinforced our attentions toward the mechanisms of different metabolic pathways, microbial interactions, accumulation of different types of metabolites, genes cascades, upregulation, and downregulation of various genes (Meena et al., 2017).

Proteomics, metabolomics, transcriptomics, and peptidomics can offer innovative ideas concerning the interaction of the plant with the environment, the internal functioning of plants, and cell-to-cell communication. RNA-Seq and gene chips are used for analysis in transcriptomics, SNP genotyping is used in genomics, ELISA, gel electrophoresis, protein microarrays, chromatography is used for further studies in proteomics, NMR, and chromatography are used in metabolomics (Gemperline, Keller, & Li, 2016). Genomic research revolves around the complete understanding of the functions of the genome at the whole genome level while proteomics aims to understand the systematic analysis of proteins. Metabolomics aims to understand the cellular status at the time of plant development (Barh, Khan, & Davies, 2015), while another field of omics named phenomics aims to understand the systematic analysis of traits in plants. During the last couple of years, phenomics has gained more progress due to the emergence of imaging techniques and the development of novel sensors for a variety of organs and traits (Brown et al., 2014; Fiorani & Schurr, 2013; Furbank & Tester, 2011).

Sequencing also facilitates the identification of genes that have been involved in various agricultural traits, for example through HTS, various studies have been made on miRNA involved in the ovule and fiber development in cotton. These studies have led to the identification of sixty-five miRNA families that are found to be conserved in cotton and from these sixty-nine families, fifty-nine miRNAs were found to express significantly. After the identification of miRNAs, computational approaches were used for target identification. A total of 1498 miRNA-target were found that comprised of 820 genes belonging to ninety-nine miRNA families. The results have shown that miR171, miR828, miR160, and miR164 are concerned with the fiber development in cotton plants. The sequencing of cotton genomes offers an extensive amount of genetic information which have not available in the past for example variants and genomic structure of the cotton (Xie, Wang, Sun, & Zhang, 2015).

15.3 Bioinformatics resources for cotton-omics

Cotton is an economically important fiber crop, due to which it attracted the attention of evolutionary biologists and taxonomists. To improve the yield and quality of cotton, understanding the genome structure and function is very important. To do this, a functional understanding of bioinformatics resources such as analysis tools, software, and databases is required (Sripathi et al., 2016). Below, we discussed the availability of bioinformatics resources (Table 15.1) for the different areas of cotton omics.

TABLE 15.1 List of bioinformatics resources used in different omics approaches to explore plant omics.

Omics approaches	Resources	Description	URL	References
Genomics	Blast2GO	Analysis and functional annotation of plant genomes	https://www.blast2go.com/	(Conesa et al., 2005)
	CottonFGD	Integrated functional genomics database	https://cottonfgd.org	(Zhu et al., 2017)
	CGRD	Database of cotton genome resources	http://cgrd.hzau.edu.cn/index.php	(Ashraf et al., 2018)
	cottonGen	Database of cotton genetics, breeding and genomic	https://www.cottongen.org/	(Yu et al., 2014)
	ccNET	Database of gene coexpression networks	http://structuralbiology.cau.edu.cn/gossypium/	(You et al., 2017)
	CottonDB	Database of cotton genome	http://www.cottondb.org	(Yu et al., 2007)
	GraP	Functional genomics studies of <i>G. raimondii</i>	http://structuralbiology.cau.edu.cn/GraP/about.html	(Zhang et al., 2015)
Transcriptomics	Trinity	Transcriptome assembler	http://TrinityRNASeq.sourceforge.net	(Pollard et al., 2009)
	TopHat	Aligner	http://tophat.cbcb.umd.edu/	(Trapnell, Pachter, & Salzberg, 2009)
	Cufflinks	Transcriptome assembler	http://cufflinks.cbcb.umd.edu/	(Trapnell et al., 2010)
	TRAPID	Transcriptome analyzer	http://bioinformatics.psb.ugent.be/webtools/trapid/	(Van Bel et al., 2013)
	EGENES	Transcriptome-based plant database	http://www.genome.jp/kegg-bin/create_kegg_menu?category=plants_egenes	(Masoudi-Nejad et al., 2007)
Proteomics	OpenMS	Analyzer for mass spectrometry data	http://www.openms.de	(Röst et al., 2016)
	SALAD	Protein sequence annotations	http://salad.dna.affrc.go.jp/salad/	(Mihara, Itoh, & Izawa, 2010)
	UniProtKb	Protein sequence information	http://www.uniprot.org/	(Boutet, Lieberherr, Tognolli, Schneider, & Bairoch, 2007)
	COGs	Classification of proteins on phylogenetic basis	https://www.ncbi.nlm.nih.gov/COG/	(Tatusov et al., 2001)
	InterPro	Collection Protein families	http://www.ebi.ac.uk/interpro/	(Mitchell et al., 2015)
	PDB	Protein structure storehouse	http://www.rcsb.org/	(Sussman et al., 1998)
	BMRB	Storehouse of NMR results of proteins	http://www.bmrwisc.edu/	(Markley et al., 2008)
	EMDB	Storehouse of Electron microscopy results of protein	https://www.ebi.ac.uk/pdbe/emdb/	(Patwardhan, 2017)
	HMMTOP	transmembrane topology	http://www.enzim.hu/hmmtop/	(Tusnady & Simon, 2001)
	ModEval	Structure evaluation	http://modbase.compbio.ucsf.edu/evaluation/	(Pieper et al., 2006)
	CAVER	Analysis and visualization tool	http://www.caver.cz/	(Chovancova et al., 2012)
	CASTp	Visualization tool	http://sts.bioe.uic.edu/	(Rayalu et al., 2012)
	ProFunc	Protein function	https://www.ebi.ac.uk/thornton-srv/databases/profunc/	(Laskowski, Watson, & Thornton, 2005)

(Continued)

TABLE 15.1 (Continued)

Omics approaches	Resources	Description	URL	References
	RADAR	Detection of repeats	https://www.ebi.ac.uk/Tools/pfa/radar/	(Heger & Holm, 2000)
	SMART	Identification of domains and motifs	http://smart.embl-heidelberg.de/	(Ponting, Schultz, Milpetz, & Bork, 1999)
	CDART	Prediction of conserved domains	https://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi	(Geer, Domrachev, Lipman, & Bryant, 2002)
	PMut	Prediction of protein mutations	http://mmb.irbbarcelona.org/PMut/	(López-Ferrando, Gazzo, De La Cruz, Orozco, & Gelpí, 2017)
	I-mutant2.0 Server	Protein stability prediction	http://folding.biofold.org/i-mutant/i-mutant2.0.html	(Capriotti & Fariselli, 2005)
	PyMol	Visualization tool	https://pymol.org/2/	(Yuan, Chan, Filipek, & Vogel, 2016)
	BLASTp	Similarity search tool	https://blast.ncbi.nlm/	(Johnson et al., 2008)
	FASTA	Similarity search tool	https://fasta.bioch.virginia.edu/fasta_www2/fasta_http://www.cgi?rm = select&pgm = fa	(Donkor, Dayie, & Adiku, 2014)
	HMMER	Detects homologs	https://toolkit.tuebingen.mpg.de/#/tools/hmmer	(Finn, Clements, & Eddy, 2011)
	ClustalOmega	Similarity search tool	https://www.ebi.ac.uk/Tools/msa/clustalo/	(Sievers & Higgins, 2014)
<i>Metabolomics</i>	MetaGeneAlyse	Metabolomics data analyzer	http://metagenealyse.mpimp-golm.mpg.de/	(Daub, Kloska, & Selbig, 2003)
	MeltDB	Metabolomics data analyzer	https://meltdb.cebitec.uni-bielefeld.de	(Neuweger et al., 2008)
	Galaxy-M	Metabolomics data analyzer	https://github.com/Viant-Metabolomics/Galaxy-M	(Davidson, Weber, Liu, Sharma-Oates, & Viant, 2016)
	XCMS	Metabolomics data analyzer	https://xcmsonline.scripps.edu/	(Tautenhahn, Patti, Rinehart, & Siuzdak, 2012)
	MetaboSearch	Metabolomics data analyzer	http://omics.georgetown.edu/metabosearch.html	(Zhou, Wang, & Resson, 2012)
	metaP-server	Pathway analyzer	http://metap.helmholtz-muenchen.de/	(Kastenmüller, Römisch-Margl, Wägele, Altmäier, & Suhre, 2011)
	MetExplore	Pathway analyzer	http://metexplore.toulouse.inra.fr/	(Cottret et al., 2010)
	MetAssign	Probabilistic annotation of metabolites	http://mzmatch.sourceforge.net/	(Daly et al., 2014)
	MetPA	Pathway analyzer	http://metpa.metabolomics.ca/	(Xia & Wishart, 2010)
	MZedDB	Interactive annotation tool	http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html	(Draper et al., 2009)
	MSEA	Pathway analyzer	http://www.metaboanalyst.ca/	(Xia & Wishart, 2010)
	ADAP	Data processing	http://www.du-lab.org/software.html	(Jiang et al., 2010)
	MMCD	Metabolomics data analyzer	http://mmcd.nmrfa.wisc.edu/	(Cui et al., 2008)
	GenePattern	Analysis and visualization tool	http://software.broadinstitute.org/cancer/software/genepattern/	(Reich et al., 2006)

15.3.1 Genomics

Genome research promises the development of genetically modified plants. Today, genome research is a rising star in the field of omics. Genomic study of cotton is far behind in comparison to other model plants (Zhang, Li, Wang, & Chee, 2008). Significant achievements made on the cotton genome are summarized below.

15.3.1.1 Translational genomics

Recent advancements in the field of genome research enable researchers to utilize genomics for developing efficient breeding strategies (Morgante & Salamini, 2003). The knowledge regarding the organization along with the structure of the genome in multiple species of plants might be capable of providing a new paradigm that how natural and artificial selection influence genes to respond better under the various environmental conditions (Beissinger et al., 2014). Through Sanger sequencing, only a few model plant genomes have been sequenced. There has been a great expectation that a large part of these sequencing studies might be transmitted to the other plant species. Occasionally, this approach is named translational genomics (Kang et al., 2016). The main objective behind translational genomics is to transfer the knowledge of the newly generated genome to the unstudied plant genome (Salentijn et al., 2007).

With the advancement in NGS technology, the knowledge of plant genomes has been increased day by day due to low sequencing costs. Translational genomics has been used extensively to study multiple traits in the plant such as plant developmental processes, stress-tolerant, etc. In polyploids, especially in the case of cotton, breeding such complex quality traits is considered to be a very difficult task (Salentijn et al., 2007). By transferring the genomic information of Arabidopsis, soybean (Schmutz et al., 2010), sorghum (Paterson et al., 2009), poplar (Tuskan et al., 2006), and rice (Goff et al., 2002; Yu et al., 2002), such complicated problems can be sorted out. The candidate-gene approach is considered a new paradigm for cotton breeding (Pflieger, Lefebvre, & Causse, 2001). Recently, it has been suggested that *Gossypium raimondii* shows more alignment with the *Vitis vinifera* and *Arabidopsis*, hence revealed that the transcriptional genomics approach can be fruitful to better understand the genome (Lin et al., 2010). It has been suggested that in case of cotton fiber, translational genomics might be helpful for making enormous improvements in variety of biomass crops. Hence this led to the identification of function of unknown gene in cotton.

15.3.1.2 Epigenomics

Several studies have been made on the mechanisms adopted by a cell to better perform its functions. As the cell comprises the same set of the gene, so why each gene behaves differently from one another? Here comes a phenomenon named epigenetics. Epigenetics is a rapidly emerging field of scientific research that emphasizes the alteration in gene activity without any change in the DNA sequence (Kang, Daines, Warren, Cowan, & Education, 2019). The word “epigenetics” actually means “over and above the genome.” Epigenetics revolves around the study of a single locus or set of loci, while *epigenomics* describes the universal study of epigenetic changes across the whole genome, such genome is called epigenome (Weinhold, 2006).

Cotton is highly genotypic-dependent due to its regeneration ability. It plays an important role as a model plant in various studies due to its genome evolution and polyploidization (Qin & Zhu, 2011). Within the genus *Gossypium*, cotton's genome offers a wide variety of information about the genomic size variation and polyploidy agronomic importance (Chen et al., 2007). The two widely cultivated species of cotton are *G. hirsutum* and *G. barbadense*. Due to high cotton yield production, *G. hirsutum* dominates the global cotton market, while in contrast, the *G. barbadense* yields high-quality cotton (Iqbal et al., 1997; Lee, Woodward, & Chen, 2007; Lovell et al., 2007). Cotton (*Gossypium*) fibers are elongated single-celled trichomes, growing in the epidermis of outer ovule integuments. After fertilization, the cotton ovule is developed into a seed coat with the division of the ovule into inner and outer integuments. Nearly 30% of outer integuments are involved in fiber cell initiation (Osabe et al., 2014). Complex genetic and epigenetic changes lead to polyploidy formation which contributes to the improvements in fiber quality traits (Kashkush, Feldman, & Levy, 2002; Shaked, Kashkush, Ozkan, Feldman, & Levy, 2001; Wang et al., 2004). Recently, up to 500 genes have been identified that are considered to be epigenetically modified among various varieties of wild and domesticated cotton. For example, wild cotton varieties may have genes that helped to better respond to drought, hence this conception makes it possible to complement genetic and epigenetic breeding systems in cotton.

There is partial genetic diversity in cotton due to successive domestication, topical polyploidization. Campbell, Williams, and Park (2009). Interestingly, DNA methylation polymorphism event is more in cotton compared to the *G. hirsutum* (Keyte, Percifield, Liu, & Wendel, 2006). During the process of breeding and domestication, studies revealed that diversity in cotton increased due to DNA methylation. Studies explored the function of DNA methylation-

regulation that contributes to the diversity of plant phenotypes and plant development. The diversity of DNA methylation in cotton was also explored by the process of methylation-sensitive amplified polymorphism that is also called MSAP. Hence, it revealed that cotton genotypes have greater DNA methylation diversity. MSAP is also used to examine the level of post-transcriptional modifications such as methylation in tissues of *G. hirsutum* (Cao et al., 2011).

Recently scientists observed that methylated genes in wild cotton are concerned with the obstruction in flowering during the hours of daylight. On the other hand, loss of methylation in the same gene in domesticated cotton leads to induction of gene expression that allowed to reach the global cotton demand. In this modern era, breeders either use CRISPR/Cas9 or chemicals to make modifications in the methylated genes. These advancements in technologies enabled the researchers to induce the targeted change in cotton epigenome to create a new cotton breed with improved characteristics (Osabe et al., 2014).

15.3.1.3 Transcriptomics

Transcriptomics is a genome-wide approach to measure the expression level of mRNA in the genome (Brady, Long, & Benfey, 2006). Transcriptomics studies have a profound impact on major aspects of biological science because it enables the researchers to analyze the variations among the gene expression of several mRNAs both qualitatively and quantitatively (Tan, Ipcho, Trengove, Oliver, & Solomon, 2009). Transcriptomics aims to measure the genetic variability during developmental processes and stress exposure in plants (Wang et al., 2009). The very first time in 1912, an epidemic occurred in cotton producer countries due to a pathogen that caused cotton leaf curl disease (CLCuD). Transcriptomics study revealed that cotton species named *Gossypium arboreum* was naturally immune against CLCuD. It has also been studied that disease-resistant genes are concerned with the transport process and may have a critical role to play in the defense mechanisms adopted by *Gossypium arboreum* against CLCuD (Naqvi et al., 2017). Another study on cotton fiber development using transcriptomics analysis suggested that genes involved in fiber elongation are associated with carbohydrates metabolism and biosynthesis of flavonoid and phenylpropanoid (Padmalatha et al., 2012). Similarly, the same study was conducted with the aim of identification of genes that showed response against stress conditions. Mainly, the genes associated with water stress are involved in the defense, regulation of gene expression along with cellular metabolisms (Park, Scheffler, Bauer, & Campbell, 2012). Upregulation in the expression level was examined in the gene associated with the vesicular trafficking and vesicle coating during the cotton fiber development (Hovav et al., 2008).

Recently, genes induced in response to *Aspergillus flavus* were identified using comparative transcriptomics analysis. A total of 732 genes were examined which shows the response to aflatoxin. All these genes are expressed differently from each other. Upregulation of gene expression encoded for helix-loop-helix (HLH) proteins and UDP glycosylation transferase were identified against aflatoxin. Moreover, another two genes encoded for 2OG and Fe(II)-dependent oxygenase superfamily were also identified that respond against toxigenic strains on *A. flavus* (Mehanathan et al., 2018). Transcriptomics is further categorized into transcriptomics of mRNAs and transcriptomics of non-mRNAs (RNomics).

15.3.1.4 Functional genomics

In the past, several studies have been conducted to perform a comparison among structural variants in the cotton genome to explore the expression level of the gene. Functional genomics is used for understanding plant biology for exploiting genomic knowledge to improve cotton breeding to reach the global cotton demand. These advancements in the genomics era led to the development of genetically modified cotton to produce a better cotton breed that is resistant to insects. But there is slow advancement regarding the improvements in cotton at the genomic level. These improvements include quality of fiber, stress-tolerant, yield, and flowering in cotton (Guo, Wang et al., 2015; Yu et al., 2016).

The whole-genome sequence of model plants promotes the effective implementation of the cotton plant that promises the consortium-based cotton genome research. NGS and in silico analysis introduce SNP in the cotton genome. These polymorphisms offer a genetic analysis of cotton. As it has been known, the plant genome contains more copy number variations, which might contribute to study the phenotypic variations. Several studies have made it clear that the gene affected by these variations has introduced significant characteristics in cotton (Ashrafi et al., 2015; Fang et al., 2017). It has also been studied that *GhARG*, *DsRed2*, and *GhCLA1* are used for highly efficient multisite genome-editing in allotetraploid cotton. In the future, CRISPR Case9 would be an effective approach for introducing multiple mutants in the cotton genome (Wang et al., 2018).

Fiber quality becomes a primary interest in the global cotton market. Multiple studies have been conducted to predict the genes that influenced the quality of fiber such as *GhExp1*, *E6*, *GA20ox*, *PIP2s*, and *GhSusA1* (Bai et al., 2014;

Harmer, Orford, & Timmis, 2002; Jiang, Guo, Zhu, Ruan, & Zhang, 2012; John & Crow, 1992; Li, Ruan et al., 2013). It is noteworthy that the quality of fiber is highly influenced by the flowering time. Several transcription factors are involved in the floral initiation that includes *MYB*, *B3*, and *MADS*. By simply targeting the genes encoding these transcription factors in cotton, the quality of fiber can be improved (Wu et al., 2015). 73% yield loss occurs under stress exposure. Functional genomics revealed that with the identification of the stress-tolerant gene, the cotton yield has been increased. Several transcription factors along with genes and physiological processes are involved that induced stress-tolerant during the period of abiotic and biotic stress. By utilizing the candidate genes, the upregulation and downregulation of gene expression can be revealed that lead to the development of cotton breed with improved traits (Guo, Shi et al., 2015; Ranjan et al., 2012).

Recent studies give a new hand to cotton fiber development by incorporating RNA interference. RNA interference assists to predict the candidate gene involved in the fiber development, stress tolerance, fiber quality, and other agronomic characteristics of the cotton plant. RNA interference application is developed rapidly because it introduces a new paradigm in cotton genomics.

Several functional genomics databases and tools are available that enable users to acquire and figure out the information at the genomic level (Ashraf et al., 2018).

15.3.2 Proteomics

Proteins are involved in various biochemical and signaling pathways, so the proteomics studies revealed the whole molecular mechanisms behind the growth, interaction, and development of plants (Mühr & Braun, 2003). Proteomics approaches for dissecting the molecular mechanisms have been studied on the model plant as in *Arabidopsis thaliana* and *Oryza sativa* to tackle multiple environmental stress conditions (Vanderschuren, Lentz, Zainuddin, & Gruijssem, 2013).

In the last 10 years, protein expression studies have been conducted by exposing various cotton tissue to various stresses. These environmental factors include drought, salinity, and pests. This protein expression analysis has led to the production of a huge data that incriminated different proteins at a particular stress level along with their effect on cellular and subcellular metabolism of the cotton plant. Many of the cotton proteomes has been released in the last few years (Du et al., 2013; Li, Zhang et al., 2013; Wang, Zheng, Gao, & Zhou, 2012; Yang, Bian, Yao, & Liu, 2008; Zhang, Yang, Zhang, & Liu, 2013). The availability of a complete genome of cotton speed up the process of protein identification using mass spectrometry.

The cotton plant has evolved a sophisticated system to better respond to various environmental conditions without any adverse effect on its growth and development (Loka & Oosterhuis, 2012). Hence, it is necessarily important to understand the mechanism underlying the development of cotton fiber to develop better cotton breeds. Recent advancement in the proteomics technologies has advanced our knowledge in perspective of cotton fiber development and stress-tolerant. In the last five years, numerous research work has been conducted that provides evidence regarding the application of proteomics in cotton fiber development and stress tolerance (Basra & Malik, 1984).

Pathogenic fungi namely *Verticillium dahlia*, *Thielaviopsis basicola*, and *Fusarium oxysporum* caused a significantly low yield of cotton (Beckman, 1966). In response to the pathogens attack, plant cells defend themselves using different response strategies such as induced responses and constitutive responses. Hence, the understanding of complete mechanisms of pathogen-plant response might contribute to the development of cotton transgenic plants for restraining cotton plant diseases. The proteomic approach is very effective to determine the pathogen-plant interaction. Several studies indicate that pathogenesis-related (PR) protein showed a great response against fungus (Wang et al., 2011). PR10 expression in *Zea mays* showed a response to *A. flavus* infection. Reading that, PR10 was responsive against *Verticillium dahlia*, *Thielaviopsis basicola*, and *Fusarium oxysporum* in the cotton plant (Chen, Brown, Rajasekaran, Damann, & Cleveland, 2006; Dowd, Wilson, & McFadden, 2004; Patil, Pierce, Phillips, Venters, & Essenberg, 2005).

Proteomics study on cotton fiber revealed that glycolysis, hydrogen peroxide, and sugar metabolism play an important role in cotton fiber development (Pang et al., 2010; Yang et al., 2008; Zhang et al., 2013). These findings provide a new perspective to combat plant diseases using proteomics approaches. Posttranslational modification is a key regulatory step in cotton fiber development (Kumar et al., 2013; Zhang & Liu, 2013). Recently TOF/MALDI TOF approaches were applied with the aim of identification of various phosphorylation-sites in differentially expressed proteins (Zhang & Liu, 2013). Based on phosphopeptide results, three enzymes named enolase (Mujer et al., 1995), UDP-L-rhamnose synthase (Pang et al., 2010), and transketolase (Gerhardt et al., 2003) were predicted as phosphorylated (Zhang & Liu, 2013). These findings suggested that enzymes concerned with carbohydrate metabolism are involved in the elongation process of cotton fiber.

Proteomics-based studies on cotton revealed that during abiotic stress, the cotton plant adopted various strategies that include tricarboxylic acid cycle, glycolysis supports, biosynthesis of ATP, photosynthesis, and biosynthesis of different defense-related proteins. In cotton, auxin, JA, ET, and BR have been discovered to form a major part of the cotton fiber development along with stress tolerance (Deeba et al., 2012; Meng et al., 2011; Wang et al., 2012; Zheng, Wang, Liu, Shu, & Zhou, 2012). Hence, the detailed knowledge of the hormones and their associated signaling pathways might encourage the understanding of biological processes in cotton.

15.3.3 Metabolomics

In 1998, the very first time, the metabolome term was originated to monitor the metabolomics complement (Oliver, Winson, Kell, & Baganz, 1998). There is a wide disparity in our understanding of the signaling and biochemical pathways in plants (Jander et al., 2004). Most of the pathways and their associated functions are still unknown. Regarding that, many types of research have been conducted but due to practical reasons, it mainly focused on the predefined questions (Fridman & Pichersky, 2005).

Metabolomics is aimed to identify metabolites in biological samples. It incorporates multiple system biology approaches including NMR and mass spectrometry. These methods provide effective data related to omics technologies because metabolites are considered as end-products in various plant cellular pathways. In plants, the most widely studied areas of metabolomics are developmental processes, response to stress, mutant and phenotype, and interaction with environments (Wolfender, Rudaz, Hae Choi, & Kyong Kim, 2013).

Cotton fiber development is an extremely complicated process that involves various pathways including metabolic and signaling pathways. Recently, GC-MS-based metabolites profiling was carried out on the ligo-lintless-2 (Li2) mutation during cotton fiber development. This study revealed multiple pathways associated with the cell elongation process, hence very helpful to understand the metabolic processes in cotton fiber elongation. Li2 mutation in cotton changed the metabolome which leads to alteration in metabolic pathways of cotton. For example, higher accumulation of tricarboxylic acid cycle-organic acids in mutant fiber signifying the high level of nitrate assimilation (Naoumkina, Hinchliffe, Turley, Bland, & Fang, 2013). No significant studies are available on cotton hence, in the future more metabolomics studies are required to understand their metabolic pathways.

15.4 Integration of multiomics data to cope with cotton plant diseases

Integration of omics is not easy work but in actual the integration of multiomics data is turned into a challenging task for the researchers. In the case of plants, climate change has introduced many environmental changes which pave the way for the infection of new diseases and insect pests in plants, hence it is needed of the hour to understand the underlying mechanism behind the pathogenesis of plant diseases. Many techniques are available but the science of omics becomes important which enables the identification of plant-microbial interaction in respect to their genotype-phenotype spectrum (Crandall, Gold, Jiménez-Gasco, Filgueiras, & Willett, 2020). The plant-microbial interaction is concerned with genetic variability among plants that have a dramatic effect on the growth of the plant. Investigation of plant defense responses is a very complicated task. Plants generally adopt multiple immune mechanisms, for example, reactive oxygen species production, enzyme synthesis, and plant cell strengthening to better respond against the attack of pathogens (Huang, Ullah, Zhou, Yi, & Zhao, 2019). Advancement in genomics, proteomics, metabolomics, and transcriptomics leads to the identification of resistant genes involved in these pathways which help select the classical cotton breed.

Significant research on plant diseases, interaction among plant and their microbiome should be considered. This can be characterized as microbial symbioses and their genes that influence the plant-microbiome interaction. Thus a more sophisticated understanding of the microbiome might be capable of reducing the chances of plant diseases which in turn leads to better cotton yield. Omics approaches have made a significant contribution to achieving this aim (López-Mondéjar, Kostovčík, Lladó, Carro, & García-Fraile, 2017).

In cotton, multiomics offers a landscape of agronomic traits such as fiber yield and quality, resistance genes, and stress tolerance, which promises to fuel the progress of cotton genetic improvement. Multiomics might be helpful in the identification of genes involved in disease pathogenesis along with their associated signaling and metabolic pathways. Before this, the unknown function of genes must be predicted and in this respect, CRISPR Cas9 is a useful approach for genome editing at multiple sites (Peng, Jones, Liu, & Zhang, 2020). The production of cotton is highly affected by the CLCuD. An epidemic of CLCuD leads to a reduction in cotton yield. The development of transgenic cotton using

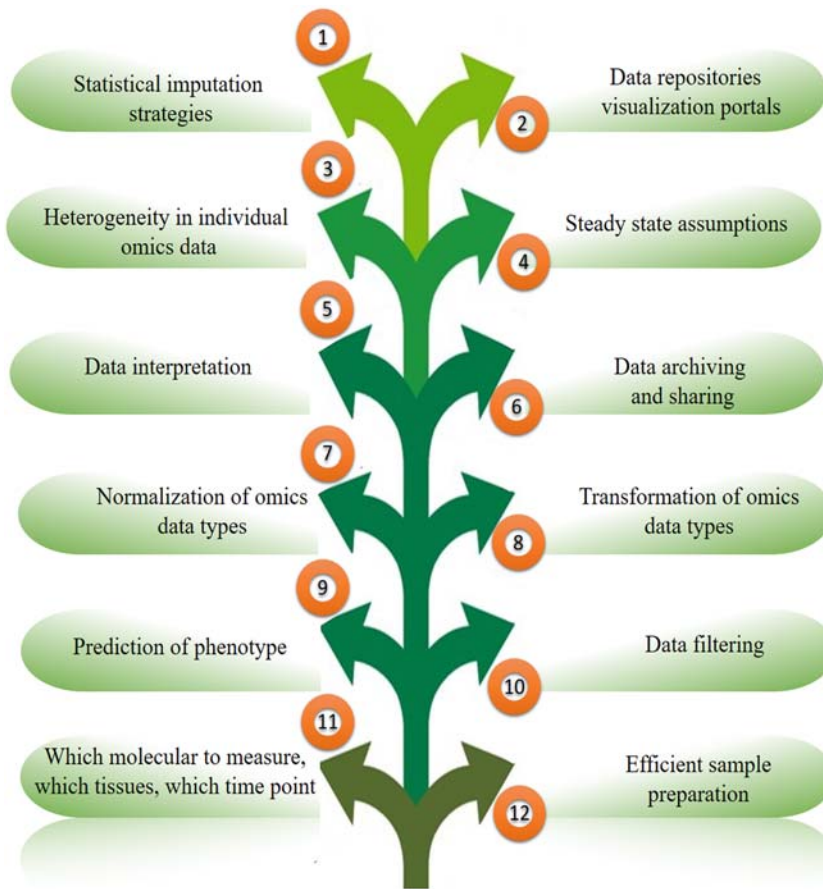


FIGURE 15.3 Overview of challenges in integrated omics.

omics approaches permitted the production of cotton plants that are resistant to the pathogens associated with the CLCuD (Rahman, Khan, Rahmat, Iqbal, & Zafar, 2017).

15.5 Challenges in the integration and analysis of multiomics data of cotton

The continuous evolution of data analysis software and databases is a compelling factor for the meaningful interpretation of multiple omics data. Over the past decades, platforms for multiomics data have greatly improved, but still, multiple challenges exist for the integration and analysis of multiomics data (Fig. 15.3) (Palsson & Zengler, 2010). The first challenge originates due to the nature of omics datasets. Omics data are very noisy and variable. Most of the omics data is qualitative which makes it very difficult to reproduce and compare (Pinu et al., 2019). While DNA data are mostly accurate and reproducible, they are mostly considered qualitative with lots of false positives based on a large number of sequences reads obtained. Datasets of proteomics, transcriptomics, and metabolomics are poorly reproducible and extremely qualitative in nature. Many of the challenges that these omics techniques have are inherent to the practices adopted by various platform users (Guo et al., 2013; Kuo, Jenssen, Butte, Ohno-Machado, & Kohane, 2002; Schloss, 2018; Sinha, Abnet, White, Knight, & Huttenhower, 2015; Tabb et al., 2010).

Lacks of appropriate metadata is another main hurdle to the effective integration of multiomics data (Hastings et al., 2019). Metadata is essential to allow the reproducibility. These are also important for the biologically relevant interpretation of the omics results. For instance, in plant multiomics analysis data about temperature trends, plant age, plant breed, watering conditions could have a substantial impact on the protein and metabolite measurements (Pinu et al., 2019).

There are many softwares and tools to integrate and analyze the multiomics data, but researchers may be unaware of all the available tools. This problem may be arising due to the lack of a central repository that links or summarize these tools. Another frequently cited problem with multiomics tools is the lack of support for achieving tool interoperability. Different bioinformatics databases have different input and output formats, several of which are nonstandard and

incompatible with the other programs. Such kinds of problems make it difficult to create a multidatabase workflow. Users may require to spend their time writing some scripts which convert and reconvert the data according to every program (Roumpeka, Wallace, Escalettes, Fotheringham, & Watson, 2017).

Statistical approaches such as correlation, clustering, and multivariate are used for multiple-omics integration, but these methods are very limited in scope and understandings into biological knowledge. For example, certain correlation approaches, such as Pearson's may be biased to outliers, while using multivariate methods, different model selection, and interpretation becomes very complicated (Cavill, Jennen, Kleinjans, & Briedé, 2016; Usadel et al., 2009).

15.6 Conclusion

Due to the advent of NGS technologies and the remarkable growth of omics data, now it looks doubtless that we are facing a big change in the era of 'big data' in biology. These technical revolutions have led to the production of a massive amount of omics data such as genomic information and production of voluminous reference genomes, RNA-sequencing for transcriptomes, and many others. Integration of multiomics data provides advantageous insights into the flow of biological info at multiple levels and consequently helps in unraveling the mechanisms underlying the biological condition of interest. Research programs that depend on the generation of multiple omics data types need to appropriately allocate resources to data processing and integration so that the full benefit of the datasets and their intrinsic information content is brought out. In part, seeking this balance is an economic challenge. Based on the challenges discussed in this study, we produced a list of recommendations that can be considered to find the potential solutions to the challenges faced during the designing of multiple-omics data integration study. These are:

- Quantitatively measure multiomics data to confirm the reproducibility and oblige the comparability
- Gather and record complete metadata to guide and notify the well-designed multiomics studies.
- Perform the power analyses, before conducting large-scale multiomics studies.
- Use quality control samples, universal standardized operating protocols, and reference standards to enable reliable multiomics measurements across the laboratories.
- Explain the clear utility of multiomics analyses to both the public and funding organizations.
- Construct centralized data repositories, reviewed database/software lists, and improved software interoperability to improve the multiomics integration.

In addition to the aforementioned recommendations, we also encourage database/software developers to take further initiative to design more user-friendly tools and software for multiomics data integration.

Acknowledgments

The work in the authors' lab is supported by Higher Education Commission and Punjab Agricultural Research Board. We apologies to those colleagues whose work could not be cited due to space constraints.

References

- Aardema, M. J., & MacGregor, J. T. (2003). *Toxicology and genetic toxicology in the new era of "toxicogenomics": Impact of "-omics" technologies*. *Toxicogenomics* (pp. 171–193). Springer.
- Afendi, F. M., Ono, N., Nakamura, Y., Nakamura, K., Darusman, L. K., Kibinge, N., et al. (2013). Data mining methods for omics and knowledge of crude medicinal plants toward big data biology. *Computational and Structural Biotechnology Journal*, 4(5), e201301010.
- Aizat, W. M., Goh, H.-H., & Baharum, S. N. (2018). *Omics Applications for Systems Biology*. Springer.
- Ashraf, J., Zuo, D., Wang, Q., Malik, W., Zhang, Y., Abid, M. A., et al. (2018). Recent insights into cotton functional genomics: Progress and future perspectives. *Plant Biotechnology Journal*, 16(3), 699–713.
- Ashrafi, H., Hulse-Kemp, A. M., Wang, F., Yang, S. S., Guan, X., Jones, D. C., et al. (2015). A long-read transcriptome assembly of cotton (*Gossypium hirsutum* L.) and intraspecific single nucleotide polymorphism discovery. *Plant Genome*, 8(2), 1–14.
- Bai, W.-Q., Xiao, Y.-H., Zhao, J., Song, S.-Q., Hu, L., Zeng, J.-Y., et al. (2014). Gibberellin overproduction promotes sucrose synthase expression and secondary cell wall deposition in cotton fibers. *PLoS ONE*, 9(5), e96537.
- Barh, D., Khan, M. S., & Davies, E. (2015). *PlantOmics: The omics of plant science*. Springer.
- Basra, A. S., & Malik, C. (1984). Development of the cotton fiber. *International Review of Cytology*, 89, 65–113.
- Beckman, C. (1966). Cell irritability and localization of vascular infections in plants. *Phytopathology*, 56(7), 821.
- Beissinger, T. M., Hirsch, C. N., Vaillancourt, B., Deshpande, S., Barry, K., Buell, C. R., et al. (2014). A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. *Genetics*, 196(3), 829–840.

- Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). *Uniprotkb/swiss-prot*. *Plant Bioinformatics* (pp. 89–112). Springer.
- Brady, S. M., Long, T. A., & Benfey, P. N. (2006). Unraveling the dynamic transcriptome. *Plant Cell*, *18*(9), 2101–2111.
- Brauer, E. K., Singh, D. K., & Popescu, S. C. (2014). *Next-generation plant science: Putting big data to work*. Springer.
- Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., et al. (2014). TraitCapture: Genomic and environment modelling of plant phenomic data. *Current Opinion in Plant Biology*, *18*, 73–79.
- Buescher, J. M., & Driggers, E. M. (2016). Integration of omics: More than the sum of its parts. *Cancer & Metabolism*, *4*(1), 1–8.
- Campbell, B., Williams, V., & Park, W. (2009). Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. *Euphytica*, *169*(3), 285–301.
- Cao, D., Gao, X., Liu, J., Kimatu, J. N., Geng, S., Wang, X., et al. (2011). Methylation sensitive amplified polymorphism (MSAP) reveals that alkali stress triggers more DNA hypomethylation levels in cotton (*Gossypium hirsutum* L.) roots than salt stress. *African Journal of Biotechnology*, *10*(82), 18971–18980.
- Capirotti, E., & Fariselli, P. (2005). Casadio R. I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, *33*(suppl_2), W306–W310.
- Cavill, R., Jennen, D., Kleinjans, J., & Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Brief Bioinformatics*, *17*(5), 891–901.
- Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T., Guo, W., et al. (2007). Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiology*, *145*(4), 1303–1310.
- Chen, Z.-Y., Brown, R., Rajasekaran, K., Damann, K., & Cleveland, T. (2006). Identification of a maize kernel pathogenesis-related protein and evidence for its involvement in resistance to *Aspergillus flavus* infection and aflatoxin production. *Phytopathology*, *96*(1), 87–95.
- Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., et al. (2012). CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Computational Biology*, *8*(10), e1002708.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*(18), 3674–3676.
- Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M.-F., et al. (2010). MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Research*, *38*(suppl_2), W132–W137.
- Crandall, S. G., Gold, K. M., Jiménez-Gasco, M. D. M., Filgueiras, C. C., & Willett, D. S. (2020). A multi-omics approach to solving problems in plant disease ecology. *PLoS ONE*, *15*(9), e0237975.
- Cui, Q., Lewis, I. A., Hegeman, A. D., Anderson, M. E., Li, J., Schulte, C. F., et al. (2008). Metabolite identification via the madison metabolomics consortium database. *Nature Biotechnology*, *26*(2), 162–164.
- Daly, R., Rogers, S., Wandy, J., Jankevics, A., Burgess, K. E., & Breitling, R. (2014). MetAssign: probabilistic annotation of metabolites from LC–MS data using a Bayesian clustering approach. *Bioinformatics (Oxford, England)*, *30*(19), 2764–2771.
- Datta, S. (2017). Advancing omics data analysis: A call for participation by a statistician in the field. *CHANCE*, *30*(2), 3026–3029.
- Daub, C. O., Kloska, S., & Selbig, J. (2003). MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics (Oxford, England)*, *19*(17), 2332–2333.
- Davidson, R. L., Weber, R. J., Liu, H., Sharma-Oates, A., & Viant, M. R. (2016). Galaxy-M: A galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *GigaScience*, *5*(1), s13742-016-0115-8.
- Deeba, F., Pandey, A. K., Ranjan, S., Mishra, A., Singh, R., Sharma, Y., et al. (2012). Physiological and proteomic responses of cotton (*Gossypium herbaceum* L.) to drought stress. *Plant Physiology and Biochemistry*, *53*, 6–18.
- Donkor, E. S., Dayie, N. T., & Adiku, T. K. (2014). Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). *Journal of Bioinformatics and Sequence Analysis*, *6*(1), 1–6.
- Dowd, C., Wilson, I. W., & McFadden, H. (2004). Gene expression profile changes in cotton root and hypocotyl tissues in response to infection with *Fusarium oxysporum* f. sp. *vasinfectum*. *Molecular Plant-Microbe Interaction*, *17*(6), 654–667.
- Draper, J., Enot, D. P., Parker, D., Beckmann, M., Snowdon, S., Lin, W., et al. (2009). Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour rules. *BMC Bioinformatics*, *10*(1), 1–16.
- Du, S.-J., Dong, C.-J., Zhang, B., Lai, T.-F., Du, X.-M., & Liu, J.-Y. (2013). Comparative proteomic analysis reveals differentially expressed proteins correlated with fuzz fiber initiation in diploid cotton (*Gossypium arboreum* L.). *Journal of Proteomics*, *82*, 113–129.
- Egan, A. N., Schlueter, J., & Spooner, D. M. (2012). *Applications of next-generation sequencing in plant biology*. Wiley Online Library.
- El-Metwally, S., Hamza, T., Zakaria, M., & Helmy, M. (2013). Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Computational Biology*, *9*(12), e1003345.
- Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nature Genetics*, *49*(7), 1089.
- Fernie, A. R., Trethewey, R. N., Krotzky, A. J., & Willmitzer, L. (2004). Metabolite profiling: From diagnostics to systems biology. *Nature Reviews Molecular Cell Biology*, *5*(9), 763–769.
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl_2), W29–W37.
- Fiorani, F., & Schurr, U. (2013). Future scenarios for plant phenotyping. *Annual Review of Plant Biology*, *64*, 267–291.
- Fridman, E., & Pichersky, E. (2005). Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. *Current Opinion in Plant Biology*, *8*(3), 242–248.
- Furbank, R. T., & Tester, M. (2011). Phenomics—technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, *16*(12), 635–644.

- Geer, L. Y., Domrachev, M., Lipman, D. J., & Bryant, S. H. (2002). CDART: protein homology by domain architecture. *Genome Research*, 12(10), 1619–1623.
- Gemperline, E., Keller, C., & Li, L. (2016). Mass spectrometry in plant-omics. *Analytical Chemistry*, 88(7), 3422–3434.
- Gerhardt, S., Echt, S., Busch, M., Freigang, J., Auerbach, G., Bader, G., et al. (2003). Structure and properties of an engineered transketolase from maize. *Plant Physiology*, 132(4), 1941–1949.
- Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science (New York, N.Y.)*, 296(5565), 92–100.
- Graw, S., Chappell, K., Washam, C. L., Gies, A., Bird, J., Robeson, M. S., et al. (2021). Multi-omics data integration considerations and study design for biological systems and disease. *Molecular Omics*.
- Greenbaum, D., Luscombe, N. M., Jansen, R., Qian, J., & Gerstein, M. (2001). Interrelating different types of genomic data, from proteome to secretome: oming in on function. *Genome Research*, 11(9), 1463–1468.
- Großkinsky, D. K., Syaifullah, S. J., & Roitsch, T. (2018). Integration of multi-omics techniques and physiological phenotyping within a holistic phenomics approach to study senescence in model and crop plants. *Journal of Experimental Botany*, 69(4), 825–844.
- Guo, J., Shi, G., Guo, X., Zhang, L., Xu, W., Wang, Y., et al. (2015). Transcriptome analysis reveals that distinct metabolic pathways operate in salt-tolerant and salt-sensitive upland cotton varieties subjected to salinity stress. *Plant Science*, 238, 33–45.
- Guo, S., Wang, Y., Sun, G., Jin, S., Zhou, T., Meng, Z., et al. (2015). Twenty years of research and application of transgenic cotton in China. *Sci Agri Sin*, 48, 3372–3387.
- Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. C., & Shyr, Y. (2013). Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS ONE*, 8(8), e71462.
- Harmer, S., Orford, S., & Timmis, J. (2002). Characterisation of six α -expansin genes in *Gossypium hirsutum* (upland cotton). *Molecular Genetics and Genomics*, 268(1), 1–9.
- Hasin, Y., Seldin, M., & Lusic, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), 1–15.
- Hastings, J., Mains, A., Virk, B., Rodriguez, N., Murdoch, S., Pearce, J., et al. (2019). Multi-omics and genome-scale modeling reveal a metabolic shift during *C. elegans* aging. *Frontiers in Molecular Biosciences*, 6, 2.
- Heger, A., & Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, 41(2), 224–237.
- Hong, H., Zhang, W., Shen, J., Su, Z., Ning, B., Han, T., et al. (2013). Critical role of bioinformatics in translating huge amounts of next-generation sequencing data into personalized medicine. *Science China Life Sciences*, 56(2), 110–118.
- Horner, D., Pavesi, G., Castrignano, T., Onorio De Meo, P. D., Liuni, S., Sammeth, M., et al. (2009). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinformatics*, 11, 181–197.
- Hovav, R., Udall, J. A., Hovav, E., Rapp, R., Flagel, L., & Wendel, J. F. (2008). A majority of cotton genes are expressed in single-celled fiber. *Planta*, 227(2), 319–329.
- Huang, H., Ullah, F., Zhou, D.-X., Yi, M., & Zhao, Y. (2019). Mechanisms of ROS regulation of plant development and stress responses. *Frontiers in Plant Science*, 10, 800.
- Iqbal M.J., Aziz N., Saeed N., Zafar Y., Malik K.J.T., Genetics A. Genetic diversity evaluation of some elite cotton varieties by RAPD analysis. 1997;94(1):139-44.
- Jabran, K., Ul-Allah, S., Chauhan, B. S., & Bakhsh, A. (2019). *An introduction to global production trends and uses, history and evolution, and genetic and biotechnological improvements in cotton*. *Cotton Production* (pp. 1–5). Wiley.
- Jander, G., Norris, S. R., Joshi, V., Fraga, M., Rugg, A., Yu, S., et al. (2004). Application of a high-throughput HPLC-MS/MS assay to *Arabidopsis* mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. *Plant J*, 39(3), 465–475.
- Jiang, W., Qiu, Y., Ni, Y., Su, M., Jia, W., & Du, X. (2010). An automated data analysis pipeline for GC – TOF – MS metabonomics studies. *J Proteome Res*, 9(11), 5974–5981.
- Jiang, Y., Guo, W., Zhu, H., Ruan, Y. L., & Zhang, T. (2012). Overexpression of GhSusA1 increases plant biomass and improves cotton fiber yield and quality. *Plant Biotechnology Journal*, 10(3), 301–312.
- John, M. E., & Crow, L. J. (1992). Gene expression in cotton (*Gossypium hirsutum* L.) fiber: cloning of the mRNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13), 5769–5773.
- Johnson, M., Zaretskaya, I., Raytselis, Y., Merezuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: A better web interface. *Nucleic Acids Research*, 36(suppl_2), W5–W9.
- Kang, J., Daines, J. R., Warren, A. N., Cowan, M. L., & Education, B. (2019). Epigenetics for the 21st-century biology student. *Journal of Microbiology & Biology Education*, 20, 3.
- Kang, Y. J., Lee, T., Lee, J., Shim, S., Jeong, H., Satyawati, D., et al. (2016). Translational genomics for plant breeding with the genome sequence explosion. *Plant Biotechnology Journal*, 14(4), 1057–1069.
- Kashkush, K., Feldman, M., & Levy, A. A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics*, 160(4), 1651–1659.
- Kastenmüller, G., Römisch-Margl, W., Wägele, B., Altmaier, E., & Suhre, K. (2011). metaP-server: a web-based metabolomics data analysis tool. *Journal of Biomedical Biotechnology*, 2011.
- Keyte, A. L., Percifield, R., Liu, B., & Wendel, J. (2006). Intraspecific D.N.A. methylation polymorphism in cotton (*Gossypium hirsutum* L.). *Journal of Heredity*, 97(5), 444–450.

- Khan, M. A., Wahid, A., Ahmad, M., Tahir, M. T., Ahmed, M., Ahmad, S., et al. (2020). *World cotton production and consumption: An overview. Cotton Production and Uses. Singapore* (pp. 1–7). Springer.
- Kobayashi M., Ohyanagi H., Yano K.J. 1 ChaPtEr Omics Databases and Gene Expression Networks in Plant Sciences. 2014:1.
- Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1), 27–38.
- Kulski, J. K. (2016). Next-generation sequencing—an overview of the history, tools, and “omic” applications. *Next generation sequencing—advances, applications and challenges*, 3–60.
- Kumar, M., Kuzhiumparambil, U., Pernice, M., Jiang, Z., & Ralph, P. J. (2016). Metabolomics: an emerging frontier of systems biology in marine macrophytes. *Algal Research*, 16, 76–92.
- Kumar, S., Kumar, K., Pandey, P., Rajamani, V., Padmalatha, K. V., Dhandapani, G., et al. (2013). Glycoproteome of elongating cotton fiber cells. *Molecular Cell Proteomics*, 12(12), 3677–3689.
- Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L., & Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics (Oxford, England)*, 18(3), 405–412.
- Laskowski, R. A., Watson, J. D., & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res*, 33 (suppl_2), W89–W93.
- Lee, J. J., Woodward, A. W., & Chen, Z. J. (2007). Gene expression changes and early events in cotton fibre development. *Ann Bot*, 100(7), 1391–1401.
- Li, D. D., Ruan, X. M., Zhang, J., Wu, Y. J., Wang, X. L., & Li, X. B. (2013). Cotton plasma membrane intrinsic protein 2s (PIP2s) selectively interact to regulate their water channel activities and are required for fibre development. *New Phytology*, 199(3), 695–707.
- Li, Y.-J., Zhang, X.-Y., Wang, F.-X., Yang, C.-L., Liu, F., Xia, G.-X., et al. (2013). A comparative proteomic analysis provides insights into pigment biosynthesis in brown color fiber. *Journal of Proteomics*, 78, 374–388.
- Lin, L., Pierce, G. J., Bowers, J. E., Estill, J. C., Compton, R. O., Rainville, L. K., et al. (2010). A draft physical map of a D-genome cotton species (*Gossypium raimondii*). *BMC Genomics*, 11(1), 395.
- Loka, D. A., & Oosterhuis, D. M. (2012). Water stress and reproductive development in cotton. In D. M. Oosterhuis (Ed.), *Flowering and fruiting in cotton The Cotton Foundation Cordova* (p. 72704). .
- López-Ferrando, V., Gazzo, A., De La Cruz, X., Orozco, M., & Gelpí, J. L. (2017). PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research*, 45(W1), W222–W228.
- López-Mondéjar, R., Kostovčík, M., Lladó, S., Carro, L., & García-Fraile, P. (2017). *Exploring the plant microbiome through multi-omics approaches: Probiotics in Agroecosystem* (pp. 233–268). Springer.
- Lovell, D., Wu, Y., White, R., Machado, A., Llewellyn, D. J., Dennis, E. S., et al. (2007). Phenotyping cotton ovule fibre initiation with spatial statistics. *Austin Journal of Botany*, 55(6), 608–617.
- Markley, J. L., Ulrich, E. L., Berman, H. M., Henrick, K., Nakamura, H., & Akutsu, H. (2008). BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *Journal of Biomolecular NMR*, 40(3), 153–155.
- Masoudi-Nejad, A., Goto, S., Jauregui, R., Ito, M., Kawashima, S., Moriya, Y., et al. (2007). EGENES: transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG. *Plant Physiology*, 144(2), 857–866.
- Medina, I., Salavert, F., Sanchez, R., de Maria, A., Alonso, R., Escobar, P., et al. (2013). Genome Maps, a new generation genome browser. *Nucleic Acids Research*, 41(W1), W41–W46.
- Meena, K. K., Sorty, A. M., Bitla, U. M., Choudhary, K., Gupta, P., Pareek, A., et al. (2017). Abiotic stress responses and microbe-mediated mitigation in plants: the omics strategies. *Frontiers in Plant Science*, 8, 172.
- Mehanathan, M., Bedre, R., Mangu, V., Rajasekaran, K., Bhatnagar, D., & Baisakh, N. (2018). Identification of candidate resistance genes of cotton against *Aspergillus flavus* infection using a comparative transcriptomics approach. *Physiology and Molecular Biology of Plants*, 24(3), 513–519.
- Meng, Y., Liu, F., Pang, C., Fan, S., Song, M., Wang, D., et al. (2011). Label-free quantitative proteomics analysis of cotton leaf response to nitric oxide. *Journal of Proteome Research*, 10(12), 5416–5432.
- Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1), 31–46.
- Mihara, M., Itoh, T., & Izawa, T. (2010). SALAD database: A motif-based database of protein annotations for plant comparative genomics. *Nucleic Acids Research*, 38(suppl_1), D835–D842.
- Mihr, C., & Braun, H.-P. (2003). *Proteomics in plant biology: Handbook of proteomic methods* (pp. 409–416). Springer.
- Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019). Integrated omics: Tools, advances and future approaches. *Journal of Molecular Endocrinology*, 62(1), R21–R45.
- Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, 43(D1), D213–D221.
- Morgante, M., & Salamini, F. (2003). From plant genomics to breeding practice. *Current Opinion in Biotechnology*, 14(2), 214–219.
- Mujer, C. V., Fox, T. C., Williams, A. S., Andrews, D. L., Kennedy, R. A., & Rumpho, M. E. (1995). Purification, properties and phosphorylation of anaerobically induced enolase in *Echinochloa phyllopogon* and *E. crus-gavonis*. *Plant Cell Physiology*, 36(8), 1459–1470.
- Naoumkina, M., Hinchliffe, D. J., Turley, R. B., Bland, J. M., & Fang, D. D. (2013). Integrated metabolomics and genomics analysis provides new insights into the fiber elongation process in Ligon lintless-2 mutant cotton (*Gossypium hirsutum* L.). *BMC Genomics*, 14(1), 155.
- Naqvi, R. Z., Zaidi, S. S.-e.-A., Akhtar, K. P., Strickler, S., Woldemariam, M., Mishra, B., et al. (2017). Transcriptomics reveals multiple resistance mechanisms against cotton leaf curl disease in a naturally immune cotton species, *Gossypium arboreum*. *Science Reports*, 7(1), 1–15.

- Neuweger, H., Albaum, S. P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., et al. (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics (Oxford, England)*, *24*(23), 2726–2732.
- Oksman-Caldentey, K.-M., & Saito, K. (2005). Integrating genomics and metabolomics for engineering plant metabolic pathways. *Current Opinion in Biotechnology*, *16*(2), 174–179.
- Oliver, S. G., Winson, M. K., Kell, D. B., & Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, *16*(9), 373–378.
- Osabe, K., Clement, J. D., Bedon, F., Pettolino, F. A., Ziolkowski, L., Llewellyn, D. J., et al. (2014). Genetic and DNA methylation changes in cotton (*Gossypium*) genotypes and tissues. *PLoS ONE*, *9*(1), e86049.
- Padmalatha, K. V., Dhandapani, G., Kanakachari, M., Kumar, S., Dass, A., Patil, D. P., et al. (2012). Genome-wide transcriptomic analysis of cotton under drought stress reveal significant down-regulation of genes and pathways involved in fibre elongation and up-regulation of defense responsive genes. *Plant Molecular Biology*, *78*(3), 223–246.
- Palsson, B., & Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nature Chemical Biology*, *6*(11), 787–789.
- Pang, C.-Y., Wang, H., Pang, Y., Xu, C., Jiao, Y., Qin, Y.-M., et al. (2010). Comparative proteomics indicates that biosynthesis of pectic precursors is important for cotton fiber and *Arabidopsis* root hair elongation. *Molecular & Cellular Proteomics*, *9*(9), 2019–2033.
- Park, W., Scheffler, B. E., Bauer, P. J., & Campbell, B. T. (2012). Genome-wide identification of differentially expressed genes under water deficit stress in upland cotton (*Gossypium hirsutum* L.). *BMC Plant Biology*, *12*(1), 1–12.
- Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, *457*(7229), 551–556.
- Patil, M. A., Pierce, M. L., Phillips, A. L., Venters, B. J., & Essenberg, M. (2005). Identification of genes up-regulated in bacterial-blight-resistant upland cotton in response to inoculation with *Xanthomonas campestris* pv. *malvacearum*. *Physiological and Molecular Plant Pathology*, *67*(6), 319–335.
- Patwardhan, A. (2017). Trends in the electron microscopy data bank (EMDB). *Acta Crystallogr D*, *73*(6), 503–508.
- Peng, R., Jones, D. C., Liu, F., & Zhang, B. (2020). From Sequencing to Genome Editing for Cotton Improvement. *Trends in Biotechnology*.
- Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of Clinical Medicine*, *9*(1), 132.
- Pflieger, S., Lefebvre, V., & Causse, M. (2001). The candidate gene approach in plant genetics: a review. *Molecular Breeding*, *7*(4), 275–291.
- Pieper, U., Eswar, N., Davis, F. P., Braberg, H., Madhusudhan, M. S., Rossi, A., et al. (2006). MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, *34*(suppl_1), D291–D295.
- Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., et al. (2019). Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites*, *9*(4), 76.
- Poisot, T., Péquin, B., & Gravel, D. (2013). High-throughput sequencing: a roadmap toward community ecology. *Ecol*, *3*(4), 1125–1139.
- Pollard, R. T., Salter, I., Sanders, R. J., Lucas, M. I., Moore, C. M., Mills, R. A., et al. (2009). Southern Ocean deep-water carbon export enhanced by natural iron fertilization. *Nature*, *457*(7229), 577–580.
- Ponting, C. P., Schultz, J., Milpetz, F., & Bork, P. (1999). SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Research*, *27*(1), 229–232.
- Qin, Y.-M., & Zhu, Y.-X. (2011). How cotton fibers elongate: a tale of linear cell-growth mode. *Current Opinion in Plant Biology*, *14*(1), 106–111.
- Rahman, M.-U., Khan, A. Q., Rahmat, Z., Iqbal, M. A., & Zafar, Y. (2017). Genetics and genomics of cotton leaf curl disease, its viral causal agents and whitefly vector: a way forward to sustain cotton fiber security. *Frontiers in Plant Science*, *8*, 1157.
- Rajasundaram, D., & Selbig, J. (2016). More effort—more results: recent advances in integrative ‘omics’ data analysis. *Current Opinion Plant Biology*, *30*, 57–61.
- Ranjan, A., Nigam, D., Asif, M. H., Singh, R., Ranjan, S., Mantri, S., et al. (2012). Genome wide expression profiling of two accession of *G. herbaceum* L. in response to drought. *BMC Genomics*, *13*(1), 94.
- Rayalu, D. J., Selvaraj, C., Singh, S. K., Ganeshan, R., Kumar, N. U., & Seshapani, P. (2012). Homology modeling, active site prediction, and targeting the anti hypertension activity through molecular docking on endothelin-B receptor domain. *Bioinformation*, *8*(2), 81.
- Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., & Mesirov, J. P. (2006). GenePattern 2.0. *Nature Genetics*, *38*(5), 500–501.
- Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weissner, H., Aicheler, F., et al. (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, *13*(9), 741–748.
- Roumpeka, D. D., Wallace, R. J., Escallettes, F., Fotheringham, I., & Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in Genetics*, *8*, 23.
- Saito, K., Hirai, M. Y., & Yonekura-Sakakibara, K. (2008). Decoding genes with coexpression networks and metabolomics—‘majority report by pre-cogs’. *Trends in Plant Science*, *13*(1), 36–43.
- Salentijn, E. M., Pereira, A., Angenent, G. C., van der Linden, C. G., Krens, F., Smulders, M. J., et al. (2007). Plant translational genomics: from model species to crops. *Molecular Breeding*, *20*(1), 1–13.
- Schaal, B. (2019). Plants and people: Our shared history and future. *Plants People Planet*, *1*(1), 14–19.
- Schatz, M. C. (2012). Computational thinking in the era of big data biology. *Genome Biology*, *13*(11), 177.
- Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio*, *9*(3).
- Schlötterer, C. (2002). A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, *160*(2), 753–763.

- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Erratum: Genome sequence of the palaeopolyploid soybean. *Nature*, 465(7294), 120.
- Shaheen, T., Iqbal, M. A., & Zafar, Y. (2016). Bioinformatics: A Way Forward to Explore “Plant Omics.” Bioinformatics-Updated Features and Applications. *IntechOpen*.
- Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., & Levy, A. A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell*, 13(8), 1749–1759.
- Sievers, F., & Higgins, D. G. (2014). Clustal omega. *Current Protocols in Bioinformatics*, 48(1), 3.13. 1–3.
- Sinha, R., Abnet, C. C., White, O., Knight, R., & Huttenhower, C. (2015). The microbiome quality control project: baseline study design and future directions. *Genome Biology*, 16(1), 1–6.
- Sirangelo, T. M. (2019). Multi-Omics Approaches in the Study of Plants. *International Journal of Advanced Research in Botany*, 5(3), 7.
- Sripathi, V. R., Buyyarapu, R., Kumpatla, S. P., Williams, A. J., Nyaku, S. T., Tilahun, Y., et al. (2016). Bioinformatics tools and genomic resources available in understanding the structure and function of *Gossypium*. *Bioinformatics (Oxford, England)*, 231.
- Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., et al. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallography D*, 54(6), 1078–1084.
- Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., et al. (2010). Repeatability and reproducibility in proteomic identifications by liquid chromatography – tandem mass spectrometry. *J Proteome Research*, 9(2), 761–776.
- Tan, K. C., Ipcho, S. V., Trengove, R. D., Oliver, R. P., & Solomon, P. S. (2009). Assessing the impact of transcriptomics, proteomics and metabolomics on fungal phytopathology. *Molecular Plant Pathology*, 10(5), 703–715.
- Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., et al. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29(1), 22–28.
- Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, 84(11), 5035–5039.
- Townsend, T. (2020). *World natural fibre production and employment. Handbook of Natural Fibres* (pp. 15–36). Elsevier.
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9), 1105–1111.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), 511–515.
- Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science (New York, N.Y.)*, 313(5793), 1596–1604.
- Tusnady, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics (Oxford, England)*, 17(9), 849–850.
- Urano, K., Maruyama, K., Ogata, Y., Morishita, Y., Takeda, M., Sakurai, N., et al. (2009). Characterization of the ABA-regulated global responses to dehydration in *Arabidopsis* by metabolomics. *Plant Journal*, 57(6), 1065–1078.
- Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., et al. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environment*, 32(12), 1633–1651.
- Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., & Vandepoele, K. (2013). TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. *Genome Biology*, 14(12), 1–10.
- Vanderschuren, H., Lentz, E., Zainuddin, I., & Gruissem, W. (2013). Proteomics of model and crop plant species: status, current limitations and strategic advances for crop improvement. *Journal of Proteomics*, 93, 5–19.
- Voora V., Larrea C., Bermudez S. **Global Market Report: Cotton. International Institute for Sustainable Development; 2020 Jun.**
- Wang, F. X., Ma, Y. P., Yang, C. L., Zhao, P. M., Yao, Y., Jian, G. L., et al. (2011). Proteomic analysis of the sea-island cotton roots infected by wilt pathogen *Verticillium dahliae*. *Proteomics*, 11(22), 4296–4309.
- Wang, J., Tian, L., Madlung, A., Lee, H.-S., Chen, M., Lee, J. J., et al. (2004). Stochastic and epigenetic changes of gene expression in *Arabidopsis* polyploids. *Genetics*, 167(4), 1961–1973.
- Wang, P., Zhang, J., Sun, L., Ma, Y., Xu, J., Liang, S., et al. (2018). High efficient multisites genome editing in allotetraploid cotton (*Gossypium hirsutum*) using CRISPR/Cas9 system. *Plant Biotechnology Journal*, 16(1), 137–150.
- Wang, X., Elling, A. A., Li, X., Li, N., Peng, Z., He, G., et al. (2009). Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell*, 21(4), 1053–1069.
- Wang, Y., Zheng, M., Gao, X., & Zhou, Z. (2012). Protein differential expression in the elongating cotton (*Gossypium hirsutum* L.) fiber under nitrogen stress. *Science China Life Sciences*, 55(11), 984–992.
- Weinhold, B. (2006). Epigenetics: the science of change. *Environmental Health Perspective*, 114, 160–167.
- Wolfender, J.-L., Marti, G., Thomas, A., & Bertrand, S. (2015). Current approaches and challenges for the metabolite profiling of complex natural extracts. *Journal of Chromatography A*, 1382, 136–164.
- Wolfender, J.-L., Rudaz, S., Hae Choi, Y., & Kyong Kim, H. (2013). Plant metabolomics: from holistic data to relevant biomarkers. *Current Medicinal Chemistry*, 20(8), 1056–1090.
- Wu, M., Li, J., Fan, S., Song, M., Pang, C., Wei, J., et al. (2015). Gene expression profiling in shoot apical meristem of *Gossypium hirsutum*. *Russian Journal of Plant Physiology*, 62(5), 684–694.
- Xia, J., & Wishart, D. S. (2010). MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics (Oxford, England)*, 26(18), 2342–2344.

- Xia, J., & Wishart, D. S. (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(suppl_2), W71–W77.
- Xie, F., Wang, Q., Sun, R., & Zhang, B. (2015). Deep sequencing reveals important roles of microRNAs in response to drought and salinity stress in cotton. *Journal Experimental Botany*, 66(3), 789–804.
- Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5), 329–342.
- Yang, Y.-W., Bian, S.-M., Yao, Y., & Liu, J.-Y. (2008). Comparative proteomic analysis provides new insights into the fiber elongating process in cotton. *Journal of Proteome Research*, 7(11), 4623–4637.
- You, Q., Xu, W., Zhang, K., Zhang, L., Yi, X., Yao, D., et al. (2017). ccNET: Database of co-expression networks with functional modules for diploid and polyploid *Gossypium*. *Nucleic Acids Research*, 45(D1), D1090–D1099.
- Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science (New York, N.Y.)*, 296(5565), 79–92.
- Yu, J., Jung, S., Cheng, C.-H., Ficklin, S. P., Lee, T., Zheng, P., et al. (2014). CottonGen: a genomics, genetics and breeding database for cotton research. *Nucleic Acids Research*, 42(D1), D1229–D1236.
- Yu J., Kohel R., Hinze L., Frelichowski J., Xu Z., Yu J., et al. Cotton DB Enhancement. 2007.
- Yu, L. H., Wu, S. J., Peng, Y. S., Liu, R. N., Chen, X., Zhao, P., et al. (2016). Arabidopsis EDT 1/HDG 11 improves drought and salt tolerance in cotton and poplar and increases cotton yield in the field. *Plant Biotechnology Journal*, 14(1), 72–84.
- Yuan, J. S., Galbraith, D. W., Dai, S. Y., Griffin, P., & Stewart, C. N., Jr (2008). Plant systems biology comes of age. *Trends in Plant Science*, 13(4), 165–171.
- Yuan, S., Chan, H. S., Filipek, S., & Vogel, H. (2016). PyMOL and Inkscape bridge the data and the data visualization. *Structure (London, England: 1993)*, 24(12), 2041–2042.
- Zaib, P., Iqbal, M., Shahzadi, R., Ahmad, H. M., Rasool, B., & Khan, S. A. (2020). Introductory Chapter: Recent Trends in “Cotton Research.” *Advances in Cotton Research. IntechOpen*.
- Zhang, B., & Liu, J.-Y. (2013). Mass spectrometric identification of in vivo phosphorylation sites of differentially expressed proteins in elongating cotton fiber cells. *PLoS ONE*, 8(3), e58758.
- Zhang, B., Yang, Y. W., Zhang, Y., & Liu, J. Y. (2013). A high-confidence reference dataset of differentially expressed proteins in elongating cotton fiber cells. *Proteomics*, 13(7), 1159–1163.
- Zhang, H.-B., Li, Y., Wang, B., & Chee, P. W. (2008). Recent advances in cotton genomics. *International Journal of Plant Genomics*, 2008.
- Zhang, L., Guo, J., You, Q., Yi, X., Ling, Y., Xu, W., et al. (2015). GraP: platform for functional genomics analysis of *Gossypium raimondii*. *Database*, 2015.
- Zheng, M., Wang, Y., Liu, K., Shu, H., & Zhou, Z. (2012). Protein expression changes during cotton fiber elongation in response to low temperature stress. *J Plant Physiology*, 169(4), 399–409.
- Zhou, B., Wang, J., & Resson, H. W. (2012). MetaboSearch: tool for mass-based metabolite identification using multiple databases. *PLoS ONE*, 7(6), e40096.
- Zhu, T., Liang, C., Meng, Z., Sun, G., Meng, Z., Guo, S., et al. (2017). CottonFGD: an integrated functional genomics database for cotton. *BMC Plant Biology*, 17(1), 1–9.

This page intentionally left blank

Omics-assisted understanding of BPH resistance in rice: current updates and future prospective

Satyabrata Nanda

MS Swaminathan School of Agriculture, Centurion University of Technology and Management, Paralakhemundi, Odisha, India

16.1 Introduction

Rice (*Oryza sativa* L.) is regarded as the most important cereal in the world, feeding more than 50% of the global population (Cheng, Zhu, & He, 2013). Consumption of rice provides more than 20% of the total calorie requirement in a typical human being. Apart from its usages as the most widely consumed staple food across the world, rice is also considered to be a model monocot plant for research. Due to the multifaceted values associated with rice, this crop has huge economic significance, especially in the countries like China, India, Vietnam, and Bangladesh. In addition, to satisfy the demands of the ever-increasing world population, there is a tremendous need of increasing the rice production. However, the global rice production is challenged by several abiotic and biotic factors. While the major abiotic factors include, salinity, drought, and temperature stress, the rice pests constitute important biotic factors. Brown planthopper (BPH) (*Nilaparvata lugens* Stål), commonly known as BPH is one of the dreadful insect pests of rice causing massive crop losses worldwide (Nanda et al., 2018). BPH is a monophagous rice pest and infests on both vegetative and reproductive stages of rice plants. It feeds on the phloem sap of rice and its excessive feeding leads to fatal drying of the plants, a phenomenon known as “hopper burn” (Cheng et al., 2013). BPH damages rice plants not only by direct feeding but also by transferring two rice viruses, including the rice grassy and the rice ragged stunt viruses. At present, the use of insecticides is the most common method of controlling BPH infestations. However, its adverse effects on environment and fear of BPH resurgence offer the need of new and more ecofriendly alternatives. Therefore controlling BPH infestations via integrated pest management (IPM) is the most promising approach.

After the advent of omics technology substantial advancement in plant science research has been attained. The rice resistance research outputs through omics-aided analysis have been proved to be crucial in discovering new insights into rice agriculture and mechanism of stress response. In the case of rice–BPH interactions, several studies have been conducted to reveal the molecular mechanisms of these complex interactions. For instance, genomics has resulted in the identification and characterization of several BPH-resistant genes (*Bph/bph*) in rice that confer rice resistance against BPH. Similarly, the use of proteomics has unearthed the role of several rice proteins in modulating the rice–BPH interactions. Also, new discoveries in the field of metabolomics have provided significant information regarding the different metabolic pathways and the rice metabolites that are involved in rice resistance response against BPH infestations. In this chapter the state-of-the-art omics technologies, namely, genomics, transcriptomics, proteomics, and metabolomics along with bioinformatics have been discussed extensively in deciphering the rice defense pathways against BPH. Moreover, a conceptual model depicting the roles of genomics, transcriptomics, proteomics, and metabolomics is also proposed related to rice resistance.

16.2 Rice genomics in brown planthopper resistance

The relatively small size of genome and diploidy nature made rice to be a perfect candidate from the cereal crops to begin the genomic studies and subsequent genome sequencing. Two rice varieties, that is, 93–11 cultivar from indica

and Nipponbare from japonica were first used for the sequencing of the rice draft genome via whole-genome-shotgun sequencing method (Goff et al., 2002; Yu et al., 2002). In the next 3 years the complete rice genomes of these two varieties were made available that covered nearly 95% of the rice genome of 389 Mb (International Rice Genome Sequencing Project, 2005; Yu et al., 2005). These studies laid the foundation of rice genomics and also served as the reference rice genome for the future resequencing studies. Upon arrival of the next-generation sequencing, many researchers did the resequencing of several rice cultivars that revealed key genetic information and strengthen rice functional genomics researches, including rice–BPH interactions (Guo, Gao, & Qian, 2014). The availability of the genome sequences and other genomic resources of rice and its wild relatives proves to be immensely beneficial in delineating rice resistance against BPH. In addition, they help in the molecular breeding of BPH-resistant rice varieties as an alternative approach under IPM. The first BPH-resistant gene *Bph1* was identified in the 1970s in Mudgo rice, and till date, 39 *Bph* genes have been identified in different rice varieties and wild rice (Zhang et al., 2020) (Table 16.1). Some of these identified genes have already been used in rice molecular breeding to produce elite rice cultivars with enhanced BPH resistance (Liu et al., 2016; Nanda et al., 2018, 2020). Moreover, genetic elements in rice conferring resistance to BPH, including both the resistance genes (*Bph*) and the quantitative trait loci (QTLs), have been identified and mapped on to different rice chromosomes (Table 16.2). The genetic mapping analysis of these identified genetic elements revealed that they are present in clusters on specific chromosomes. Chromosomes 3, 4, 6, and 12 harbor most of such clusters and are crucial for regulating the genetics of rice–BPH-resistant response (Cheng et al., 2013). For instance, 12 BPH-resistant genes are clustered on chromosome 4, while 8 of them are present on chromosome 12. Apart from the BPH-resistant genes, some QTLs have also been identified by using different crossing methods and from different mapping populations. For example, qBPH6(t) was mapped onto chromosome 6 lying between a set of microsatellite markers RM469 and RM568 (Jairin et al., 2007). Similarly, qBPH3 was mapped onto chromosome 3, while qBPH4, qBPH4.2, qBPH4.3, and qBPH4.4 were mapped onto chromosome 4 (Hu et al., 2015; Mohanty et al., 2017). It is hypothesized that the cluster of genetic elements could be the tightly linked genes or different alleles of a particular gene or genes having multiple insect-response specificities (Du, Chen, Guo, & He, 2020). The eight BPH-resistant genes clustered on chromosome 12 were isolated and analyzed to be different alleles of the same gene (Zhao et al., 2016). These small allelic variations result in great variation in the rice resistance levels against BPH infestations. Several studies have revealed that the BPH feedings on rice phloem sap, production of honeydew, fecundity rate, and mortality rate goes down when BPH infest on the rice varieties carrying these *Bph* genes (Cheng et al., 2013). Thus multiple efforts have been made to use the *Bph* genes in molecular breeding of rice resistance through marker-assisted selection and gene pyramiding (Hu et al., 2012, 2013; Qiu, Guo, Jing, Zhu, & He, 2012). These improved rice germplasms offer a better tolerance/resistance capacity to BPH infestations and act as the source for the development of elite rice varieties.

On the other hand, mere identification of the rice genetic elements conferring resistance to BPH pressure is not enough. Those genes need to be isolated and characterized. The structural and functional characterization of such genes will be of huge interest in rice molecular breeding against BPH infestations and understanding the underlying molecular mechanisms of rice–BPH interactions. In this connection, *Bph14* was the first gene to be cloned and characterized in rice from chromosome 3 (Du et al., 2009). Structural analysis of *Bph14* revealed it to be from the NLR family of resistance gene with a CC-NB-LRR conserved domain. Functional genomics studies confirmed that *Bph14* modulates rice resistance against BPH by activating the salicylic acid (SA) signaling and increased callose depositions at rice sieve tubes (Du et al., 2009). The SA signaling ensures an improved resistance response at a molecular level, much like system acquired resistance. On the other hand, the callose depositions make the sieve tubes difficult to penetrate for the BPH, resulting in reduced growth and survival of BPH. In addition, the functional genomics studies on *Bph* genes revealed that rice carries multiple NLR family proteins that confer BPH resistance. For instance, Bph26 was found to be another NLR protein having a CC-NBS-LRR domain (Tamura et al., 2015). Interestingly, *Bph18* and *Bph26* were identified to share the same locus on rice chromosome. Bph18 is also an NLR protein containing a dual nucleotide binding (CC-NB-NB-LRR) domain (Ji et al., 2016a). Similarly, *Bph9* was cloned and characterized to be a CC-NB-NB-LRR resistance gene conferring both antibiosis and antixenosis against BPH (Zhao et al., 2016). In addition, molecular cloning and characterization of *Bph6* revealed that it is an LRR-type resistance protein, localized in the exocyst (Guo et al., 2018). Under BPH attacks, elevated expressions of *Bph6* activate downstream signaling cascades, including SA and jasmonic acid (JA) pathways, facilitate exocytosis, and induce the cell wall reinforcements. Conversely, another widely explored BPH-resistant gene in rice, that is, *Bph3* is a cluster of lectin receptor–like kinases that is localized at the cell membrane. Cloning and characterization of *Bph3* resulted in the identification of the members of this cluster to be *OsLecRK1–4* in the Rathu Heenati cultivar (Liu et al., 2015). Its functional validation revealed that *OsLecRKs* are involved in rice–BPH interactions and provide a wide-spectrum rice resistance against insect pests (Liu et al., 2015; Nanda et al., 2018). Overall, these diversities of the BPH-resistant genes/QTLs in rice offer a better opportunity for successful BPH resistance.

TABLE 16.1 List of brown planthopper (BPH)-resistant (*Bph*) genes identified in rice varieties and its wild relatives.

Gene name	Source of identification	References
<i>Bph1</i>	Mudgo	Athwal, Pathak, Bacalangco, and Pura (1971)
<i>bph2</i>	ASD7	Athwal et al. (1971)
<i>Bph3</i>	Rathu Heenati	Lakshminarayana and Khush (1977), Liu et al. (2015)
<i>bph4</i>	Babawee	Sidhu and Khush (1978)
<i>bph5</i>	ARC 10550	Khush, Karim, and Angeles (1985)
<i>Bph6</i>	Swarnalata	Kabis and Khush (1988)
<i>bph7</i>	T12	Kabis and Khush (1988)
<i>bph8</i>	Chin Saba	Nemoto, Ikeda, and Kaneda (1989)
<i>Bph9</i>	Pokkali	Nemoto et al. (1989)
<i>Bph10</i>	<i>O. australiensis</i>	Ishii, Brar, Multani, and Khush (1994)
<i>bph11</i>	<i>O. officinalis</i>	Hirabayashi (1998)
<i>Bph12</i>	B14 (<i>O. officinalis</i>)	Qiu et al. (2012)
<i>Bph13</i>	<i>O. officinalis</i>	Renganayaki et al. (2002)
<i>Bph14</i>	B5 (<i>O. officinalis</i>)	Du et al. (2009)
<i>Bph15</i>	B5 (<i>O. officinalis</i>)	Yang et al. (2004)
<i>Bph16</i>	M1635–7	Hirabayashi et al. (2004)
<i>Bph17</i>	Rathu Heenati	Sun, Su, Wang, Zhai, and Wan (2005)
<i>Bph18</i>	<i>O. australiensis</i>	Ji et al. (2016b)
<i>bph18(t)</i>	IR65482–7-216–1-2	Jena, Jeung, Lee, Choi, and Brar (2006)
<i>bph19(t)</i>	AS20–1	Chen, Wang, Pang, and Pan (2006)
<i>Bph20(t)</i>	IR71033–121–15	Rahman et al. (2009)
<i>Bph21(t)</i>	IR71033–121–15	Rahman et al. (2009)
<i>Bph22(t)</i>	<i>O. glaberrima</i>	Ram et al. (2010)
<i>Bph23(t)</i>	<i>O. minuta</i>	Ram et al. (2010)
<i>bph24(t)</i>	IR72678–6–9-B	Deen and Rammesh (2008)
<i>Bph25</i>	ADR52	Myint et al. (2012)
<i>Bph26</i>	ADR52	Tamura et al. (2015)
<i>Bph27(t)</i>	Balamawee	He et al. (2013)
<i>Bph28(t)</i>	DV85	Wu et al. (2014)
<i>bph29</i>	RBPH54	Wang et al. (2015)
<i>Bph30</i>	AC-1613	Wang et al. (2018)
<i>Bph31</i>	CR2711–76	Prahalada et al. (2017)
<i>Bph32</i>	PTB33	Ren et al. (2016)
<i>Bph33</i>	KOLAYAL	Hu, Chang, Zou, Tang, and Wu (2018)
<i>Bph34</i>	IRGC104646	Kumar et al. (2018)
<i>Bph35</i>	RBPH660	Zhang et al. (2020)
<i>Bph36</i>	RBPH16, RBPH17	Li et al. (2019)
<i>Bph37</i>	IR64	Yang et al. (2019)
<i>Bph38(t)</i>	BC ₁ F ₅ of HHZ x Khazar	Balachiranjeevi et al. (2019)
<i>bph39(t)</i>	RPBio4918–230S	Akanksha et al. (2019)
<i>Bph40(t)</i>	RPBio4918–230S	Akanksha et al. (2019)

TABLE 16.2 Chromosome location of different *Bph* genes identified in rice.

Gene name	Chromosome	References
<i>Bph1</i>	12L	Hirabayashi (1998)
<i>bph2</i>	12L	Murai et al. (2001)
<i>Bph3</i>	4S	Lakshminarayana and Khush (1977)
<i>bph4</i>	6S	Jairin et al. (2007)
<i>Bph6</i>	4L	Kabis and Khush (1988)
<i>bph7</i>	12L	Kabis and Khush (1988)
<i>Bph9</i>	12L	Su et al. (2006)
<i>Bph10</i>	12L	Lang and Buu (2003)
<i>bph11</i>	3L	Hirabayashi (1998)
<i>Bph12</i>	4L	Qiu et al. (2012)
<i>Bph13</i>	3S	Renganayaki et al. (2002)
<i>Bph14</i>	3L	Du et al. (2009)
<i>Bph15</i>	4S	Yang et al. (2004)
<i>Bph17</i>	4S	Sun et al. (2005)
<i>Bph18</i>	12L	Jena et al. (2006)
<i>bph19(t)</i>	3S	Chen et al. (2006)
<i>Bph20(t)</i>	4S	Rahman et al. (2009)
<i>Bph21(t)</i>	12	Rahman et al. (2009)
<i>Bph22(t)</i>	6S	Harini et al. (2010)
<i>Bph25</i>	6S	Myint et al. (2012)
<i>Bph26</i>	12L	Tamura et al. (2015)
<i>Bph27(t)</i>	4L	He et al. (2013)
<i>Bph28(t)</i>	11L	Wu et al. (2014)
<i>bph29</i>	6S	Wang et al. (2015)
<i>Bph30</i>	4S	Wang et al. (2018)
<i>Bph31</i>	3L	Prahalada et al. (2017)
<i>Bph32</i>	6S	Ren et al. (2016)
<i>Bph33</i>	4S	Hu et al. (2018)
<i>Bph34</i>	4L	Kumar et al. (2018)
<i>Bph36</i>	4S	Li et al. (2019)
<i>Bph37</i>	1	Yang et al. (2019)
<i>Bph38(t)</i>	1L	Balachiranjeevi et al. (2019)

16.3 Rice transcriptomics in brown planthopper resistance

Analysis of plant transcriptomes under stress conditions has provided new insights by revealing more details about plant stress biology. It has proven to be effective in delineating the complexes of plant–insect interactions. In rice, several transcriptome studies have been performed to get better insights into rice–BPH interactions (Jing et al., 2017; Kumar et al., 2020; Tan et al., 2020). Prior to the transcriptome era, the identification of rice differentially expressed genes (DEGs) in response to BPH attacks was performed by using cDNA-amplified fragment length polymorphism (cDNA-AFLP). By the use of cDNA-AFLP, several

rice transcripts were identified to be involved in physiological processes, including stress signaling, transcriptional control, and detoxification (Yang, Zhang, Zhu, & He, 2006). Further, Zhang, Zhu, and He (2004) reported multiple cDNA clones to be upregulated in the resistant B5 rice variety in response to insect feeding and pathogen infection. These cDNAs were functionally attributed to major pathways, including stress signaling, cellular oxidation, wound response, and pathogen-related protein synthesis. After the advent of the next-generation sequencing and transcriptome profiling, the rice transcript dynamics under BPH infestations were better explored. Gene expression profiles in a BPH-resistant rice carrying the *Bph15*, exhibited constitutive expression of the LRR domain-containing gene and jacalin-related lectin protein genes (Lv et al., 2014). The comparative transcriptome analysis of two rice varieties, one susceptible and the other being resistant to BPH, revealed that several genes related to SA and JA signaling were upregulated in the resistant rice (Li et al., 2017). Similarly, differential expression analysis in rice revealed that BPH infestation caused the upregulation of lignin biosynthesis and antioxidant genes that aid in the rice resistance (Jannoey, Channei, Kotcharerk, Pongprasert, & Nomura, 2017). In addition, BPH feedings activated the wound-induced genes and signaling responses, including the mitogen-activated protein kinase (MAPK) signaling cascade. Multiple MAPKs have been reported to be involved in the rice–BPH interactions (Nanda et al., 2018). For instance, OsMPK3 that regulated JA, jasmonoyl-L-isoleucine (JA-Ile), and SA levels in rice under BPH infestation plays crucial role in rice–BPH interactions (Zhou et al., 2019).

Apart from the mRNA transcripts of the key genes, several small noncoding RNA take part in the rice–BPH interactions. Small RNA transcriptome or sRNA profiling of rice varieties under BPH infestations revealed that microRNA (miRNA) plays a major role in modulating the rice resistance pathways. In a recent study the combined transcriptome and miRNA analysis revealed that 34 DE miRNAs targeted 42 DEGs in rice and these mRNA–miRNA pairs served as candidates for BPH resistance in rice (Tan et al., 2020). In addition to that, multiple miRNAs were found to show differential expressions during compatible and incompatible rice–BPH interactions. For instance, 104 DE miRNAs were identified in a *Bph15* introgression line as compared to the susceptible recipient rice plant (Wu et al., 2017). Similarly, in another study, 138 numbers of DE miRNAs were found during the compatible rice–BPH interaction, whereas 140 DE miRNAs were obtained during the incompatible interaction (Nanda et al., 2020). Furthermore, two miRNAs, including OsmiR156 and OsmiR396, have been considered as the primary regulators of the rice–BPH interactions (Nanda, Mishra, & Joshi, 2021). Both of these miRNAs negatively modulate the rice resistance against BPH; OsmiR156 modulate the rice JA pathway, whereas OsmiR396 regulates the flavanone 3-hydroxylase that in turn controls the flavonoid biosynthetic pathway (Ge et al., 2018; Dai et al., 2019). These findings suggest that analysis of rice transcriptome and small RNA profiling provides an illustrative idea of the rice resistance response against BPH. Moreover, analysis of these mRNA and miRNA profiles is hugely beneficial in understanding the transcriptional and posttranscriptional regulations of several genes of the major pathways linked to rice immune response.

16.4 Rice proteomics in brown planthopper resistance

Proteomics have gained enormous importance recently for their ability to produce better insights into the cellular physiological changes, including plant stress responses. Although huge data have been generated based on proteomic studies of rice against several abiotic and biotic stresses, under BPH infestation the rice comparative proteomics still offers scope to explore. To explore rice proteome, techniques like mass spectrometry (MS) and isobaric tag-based methodology for relative peptide quantification (iTRAQ) are the most preferred ones. iTRAQ often works coupled with the multidimensional liquid chromatography and MS enabling the accurate assessment of protein levels in rice samples (Hussain et al., 2019). These proteomic studies have proven to be a powerful approach to unravel the complex rice molecular responses against BPH (Agarwal et al., 2016). For example, the comparative proteome analysis of the *Bph15* containing resistance rice and susceptible Taichung Native 1 (TN1) rice revealed that the upregulation of the glycine cleavage system H-protein could be an important and unique rice defense system against BPH infestation (Wei et al., 2009). In another study the differential proteome analysis between IR64 rice and its two near-isogenic mutant varieties was performed to evaluate the relative expression of proteins during rice–BPH interaction (Sangha et al., 2013). The results revealed that the BPH-resistant mutant lines possessed elevated proteins related to stress response and protein synthesis metabolism as compared to the wild types. Similarly, upregulation of several proteins, including a heat shock protein (HSP20), two lipoxygenases (LOX), an Ent-cassa-12,15-diene synthase (DTC1), and two dirigent proteins (DIRs) with potential roles in BPH-resistant breeding, was observed in the BPH-resistant lines originated from *Oryza officinalis* under BPH pressure (Zhang et al., 2019). As BPH is a sucking pest of rice feeding exclusively on phloem sap, both the salivary gland of BPH and the phloem exudates become valuable resources to understand the molecular mechanism of rice–BPH interactions. Many studies have been carried out to understand the protein dynamics and to identify candidate effectors in BPH salivary repertoire. On the other hand, the comparative proteome analysis of rice

phloem exudates of the resistant *Bph14* and *Bph15* introgressed lines and the susceptible 9311 recipient variety revealed that proteins specific for defense signaling, cellular proteins, and carbohydrate metabolism were accumulated in the resistant plants (Du et al., 2015). These comparative proteome studies have not only helped to understand the molecular dynamics of BPH infestation in rice but also being extrapolated to unravel the interactions of rice and small BPH (SBPH). Moreover, advancements in the field of proteomics and comparative DEP analysis have greatly facilitated the understanding of rice defense responses against BPH attacks.

16.5 Rice metabolomics in brown planthopper resistance

The analysis of plant metabolites and complete metabolome reveals several key details about the plant physiological processes, especially in stress response. Infestation of BPH on rice drives series of metabolic activities that in turn can modulate the rice resistance responses (Cheng et al., 2013). Depending on the type of rice–BPH interaction, that is, compatible or incompatible, the rice metabolite repertoire varies accordingly. The evolutionary developments in the omics led to the popularization of rice metabolomics studies under BPH pressure by employing techniques, including gas chromatography (GC), nuclear magnetic resonance, and MS (Kang, Yue, Xia, Liu, & Zhang, 2019). Moreover, rice metabolomics studies are combined with the transcriptomic or gene expression studies to derive better and insightful inference on rice–BPH interactions. For instance, the combined gene expression and metabolomics study in B5 rice, a resistant variety carrying *Bph14* and *Bph15*, revealed that BPH feeding reprogrammed the rice metabolome as compared to the susceptible TN1 rice. Most of the metabolites with elevated concentrations were found to be involved in processes like carbohydrate metabolism, amino acid metabolism, and other stress-responsive secondary metabolisms (Liu et al., 2010). Upregulation of carbohydrate metabolism and metabolism of amino/nucleotide sugars were observed in the resistance Mudgo rice under BPH attacks (Ji et al., 2013). Similarly, another study involving the GC–MS-mediated metabolomic analysis of rice leaf sheaths reported that BPH infestations induce glycolysis, β -oxidation, and shikimic acid synthesis pathway during the incompatible rice–BPH interactions (Peng et al., 2016). On the other hand, the comparative metabolomic study between two BPH-resistant rice varieties IR56 and IR36 showed that in IR56 rice the thiamine, taurine, and hypotaurine metabolisms were increased, whereas in IR36 rice the cyano-amino acids and lipids metabolism were abundant (Kang et al., 2019). In addition, the metabolic profiling of the resistant rice IL308 and susceptible rice variety KDML105 revealed that the levels of apigenin 6-C- α -L-arabinoside-8-C- β -L-arabinoside, rhoifolin, schaftoside, and iso-schaftoside were comparatively higher in the resistant plant than the susceptible under BPH infestation (Uawisetwathana et al., 2019). All these findings indicate that the metabolite pool in rice has a significant role in determining the success of rice immune response against BPH attacks. Moreover, a complex yet fine-tuned network among the rice genomics, transcriptomics, proteomics, and metabolomics could be responsible for regulating the rice–BPH interactions and their molecular consequences (Fig. 16.1).

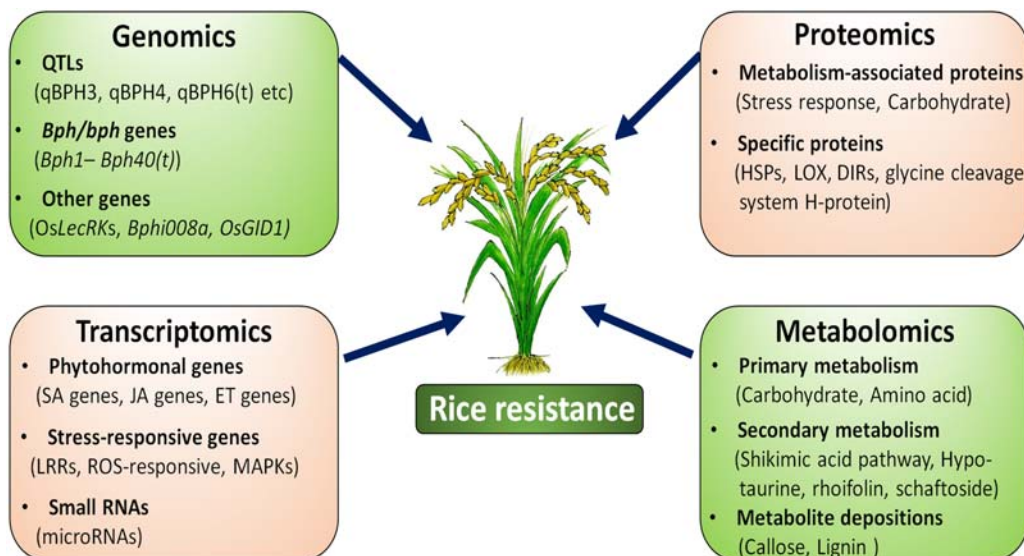


FIGURE 16.1 A schematic representing the holistic approach for BPH resistance in rice by using the omics approaches. *BPH*, Brown planthopper.

16.6 Bioinformatics in brown planthopper resistance in rice

Bioinformatics plays an important role in the modern omics-driven researches. In silico analysis of genes and prediction of different interaction models have proven to be an effective mean of the functional genomics research. In rice–BPH interactions, several rice genetic elements, including QTL and *Bph* genes, have been identified by the computational biology applications. For instance, using bioinformatics and functional analysis approaches, *Bph32* was identified in the PTB33 rice (Ren et al., 2016). Further structural analysis revealed that the gene encodes a protein containing a short consensus repeat. In addition to this, advances in bioinformatics have led to the establishment of several dedicated databases and tools that provide vital information on rice genes and proteins. For instance, databases like PhosphoRice, PRIN, and Oryza PG-DB have significantly contributed to rice resistance researches. PhosphoRice is a *meta*-predictor of rice-specific phosphorylation sites, whereas PRIN predicts the rice interactome network (Gu, Zhu, Jiao, Meng, & Chen, 2011; Que et al., 2012). On the other hand, Oryza PG-DB serves as a rice proteome database store and disseminates information on shotgun proteogenomics (Helmy, Tomita, & Ishihama, 2011). Further, the use of computational biology and different bioinformatics tools are common in analyzing the sRNA-mediated rice resistance. In particular, analysis of sRNA structure and prediction of their targets are done by using the bioinformatics tools. Recently, 246 miRNAs were identified in a *Bph3* carrying rice variety (IR56) in response to BPH feedings. In addition, analysis of differential expressive miRNA and the prediction of their targets were facilitated by bioinformatic approaches (Nanda et al., 2020). On the other hand, bioinformatic-mediated identification of candidate genes and gene families has been reported in BPH. Wan et al. (2016) performed the genome-wide identification of the basic helix-loop-helix transcription factors in BPH. Similarly, identification of the Ca²⁺/calmodulin-dependent protein kinase II isoforms in BPH was carried out by using various bioinformatic approaches (Wang, Lai, Wan, Fu, & Zhu, 2019). Moreover, the characterization of genes and their interaction models in both rice and BPH provided valuable information. Further, these in silico analysis results serve as the foundation for the wet lab validations and functional genomic studies. In addition, the databases dedicated to rice and BPH research act as the repertoire for various genes, transcription factors, and proteins.

16.7 Conclusion and future prospective

Rice and BPH interactions are complex and involve many layers, from genes to metabolites. The current understandings of these interactions have suggested that rice resistance response can be regulated by multiple players, either independently or in collaboration. On the one hand, rice multiple genetic elements, including QTLs and *Bph* genes, have been identified and mapped that provide BPH resistance. On the other hand, rapid evolution of different BPH populations can adapt to different rice varieties and cause their resistance breakdown. Therefore it's essential to look for alternative options in contributing toward rice resistance other than the resistant *Bph* genes. Apart from these genetic elements, other broad-spectral genes, including *OsLecRKs*, *OsMPKs*, and *OsGIDs*, have been reported to regulate the rice resistance against BPH. However, their applicability in real-life BPH problems is yet to be realized. Thus evaluation of their effects in fields and in different weather conditions can reveal further details. On the contrary, transcriptomic studies have provided even deeper insights into the rice transcriptional dynamics in response to BPH feedings. These results have also paved way to select and validate new candidate genes in rice–BPH interactions. Moreover, the outputs related to the roles of small RNAs are interesting, showing promising results in conferring rice resistance. The sRNA and their target gene network should be explored in detail to get even better understating of their molecular mechanisms in controlling rice resistance against BPH. Similarly, proteomics and metabolomics studies have unraveled the rice proteome and metabolome and the candidates playing roles in rice–BPH interactions. However, in-depth analysis of these candidate proteins or metabolites, such as loss of function and complementation, can provide their detailed roles. The hot-trend of CRISPR-mediated genome engineering has not well-explored in rice–BPH interactions. Therefore there is a huge scope of the application of genome editing technologies to understand the molecular mechanisms of rice–BPH interactions. Moreover, the combinational outputs of all these omics-assisted domains can better illustrate the rice–BPH interactions.

References

- Agarwal, P., Parida, S. K., Raghuvanshi, S., Kapoor, S., Khurana, P., Khurana, J. P., & Tyagi, A. K. (2016). Rice improvement through genome-based functional analysis and molecular breeding in India. *Rice*, 9, 1. Available from <https://doi.org/10.1186/s12284-015-0073-2>.
- Akanksha, S., Lakshmi, V. J., Singh, A. K., Deepthi, Y., Chirutkar, P. M., Ramdeen., ... Ram, T. (2019). Genetics of novel brown planthopper *Nilaparvata lugens* (Stål) resistance genes in derived introgression lines from the interspecific cross *O. sativa* var. Swarna × *O. nivara*. *Journal of Genetics*, 98, 113. Available from <https://doi.org/10.1007/s12041-019-1158-2>.

- Athwal, D. S., Pathak, M. D., Bacalangco, E. H., & Pura, C. D. (1971). Genetics of resistance to brown planthoppers and green leafhoppers in *Oryza sativa* L. 1. *Crop Science*, 11, 747–750. Available from <https://doi.org/10.2135/cropsci1971.0011183X001100050043x>.
- Balachiranjeevi, C. H., Prahalada, G. D., Mahender, A., Jamaloddin, M., Sevilla, M. A. L., Marfori-Nazarea, C. M., ... Ali, J. (2019). Identification of a novel locus, *BPH38(t)*, conferring resistance to brown planthopper (*Nilaparvata lugens* Stal.) using early backcross population in rice (*Oryza sativa* L.). *Euphytica*, 215, 185. Available from <https://doi.org/10.1007/s10681-019-2506-2>.
- Chen, J. W., Wang, L., Pang, X. F., & Pan, Q. H. (2006). Genetic analysis and fine mapping of a rice brown planthopper (*Nilaparvata lugens* Stål) resistance gene *bph19(t)*. *Molecular Genetics and Genomics: MGG*, 275, 321–329. Available from <https://doi.org/10.1007/s00438-005-0088-2>.
- Cheng, X., Zhu, L., & He, G. (2013). Towards understanding of molecular interactions between rice and the brown planthopper. *Molecular Plant*, 6, 621–634. Available from <https://doi.org/10.1093/mp/sst030>.
- Dai, Z., Tan, J., Zhou, C., Yang, X., Yang, F., Zhang, S., ... Shi, Z. (2019). The OsmiR396–OsGRF8–OsF3H-flavonoid pathway mediates resistance to the brown planthopper in rice (*Oryza sativa*). *Plant Biotechnology Journal*, 17, 1657–1669. Available from <https://doi.org/10.1111/pbi.13091>.
- Deen, R., & Rammesh, K. (2008). Identification of new gene for BPH resistance introgressed from *O. rufipogon*. *Genetics*, 22, 24–25.
- Du, B., Chen, R., Guo, J., & He, G. (2020). Current understanding of the genomic, genetic, and molecular control of insect resistance in rice. *Molecular Breeding*, 40, 24. Available from <https://doi.org/10.1007/s11032-020-1103-3>.
- Du, B., Wei, Z., Wang, Z., Wang, X., Peng, X., Du, B., ... He, G. (2015). Phloem-exudate proteome analysis of response to insect brown planthopper in rice. *Journal of Plant Physiology*, 183, 13–22. Available from <https://doi.org/10.1016/j.jplph.2015.03.020>.
- Du, B., Zhang, W., Liu, B., Hu, J., Wei, Z., Shi, Z., ... He, G. (2009). Identification and characterization of *Bph14*, a gene conferring resistance to brown planthopper in rice. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 22163–22168. Available from <https://doi.org/10.1073/pnas.0912139106>.
- Ge, Y., Han, J., Zhou, G., Xu, Y., Ding, Y., Shi, M., Guo, C., & Wu, G. (2018). Silencing of miR156 confers enhanced resistance to brown planthopper in rice. *Planta*, 248, 813–826. Available from <https://doi.org/10.1007/s00425-018-2942-6>.
- Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., ... Briggs, S. (2002). A Draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science (New York, N.Y.)*, 296, 92–100. Available from <https://doi.org/10.1126/science.1068275>.
- Gu, H., Zhu, P., Jiao, Y., Meng, Y., & Chen, M. (2011). PRIN: A predicted rice interactome network. *BMC Bioinformatics*, 12, 161. Available from <https://doi.org/10.1186/1471-2105-12-161>.
- Guo, J., Xu, C., Wu, D., Zhao, Y., Qiu, Y., Wang, X., ... He, G. (2018). *Bph6* encodes an exocyst-localized protein and confers broad resistance to planthoppers in rice. *Nature Genetics*, 50, 297–306. Available from <https://doi.org/10.1038/s41588-018-0039-6>.
- Guo, L., Gao, Z., & Qian, Q. (2014). Application of resequencing to rice genomics, functional genomics and evolutionary analysis. *Rice*, 7, 4. Available from <https://doi.org/10.1186/s12284-014-0004-7>.
- Harini, A. S., Lakshmi, S. S., Kumar, S. S., Sivaramkrishnan, S., & Kadirvel, P. (2010). Validation and fine-mapping of genetic locus associated with resistance to brown plant hopper [*Nilaparvata lugens* (Stal.)] in rice (*Oryza sativa* L.). *Asian Journal of Bio Science*, 5(1), 32–37.
- He, J., Liu, Y., Liu, Y., Jiang, L., Wu, H., Kang, H., ... Wan, J. (2013). High-resolution mapping of brown planthopper (BPH) resistance gene *Bph27(t)* in rice (*Oryza sativa* L.). *Molecular Breeding*, 31, 549–557. Available from <https://doi.org/10.1007/s11032-012-9814-8>.
- Helmy, M., Tomita, M., & Ishihama, Y. (2011). OryzaPG-DB: Rice proteome database based on shotgun proteogenomics. *BMC Plant Biology*, 11, 63. Available from <https://doi.org/10.1186/1471-2229-11-63>.
- Hirabayashi, H. (1998). Identification of brown planthopper resistance gene derived from *O. officinalis* using molecular markers in rice. *Breeding Science*, 48, 82.
- Hirabayashi, H., Ideta, O., Sato, H., Takeuchi, Y., Ando, I., Nemoto, H., ... Ogawa, T. (2004). Identification of a resistance gene to brown planthopper derived from *Oryza minuta* in rice. *Breeding Research*, 6, 285.
- Hu, J., Chang, X., Zou, L., Tang, W., & Wu, W. (2018). Identification and fine mapping of *Bph33*, a new brown planthopper resistance gene in rice (*Oryza sativa* L.). *Rice*, 11, 55. Available from <https://doi.org/10.1186/s12284-018-0249-7>.
- Hu, J., Cheng, M., Gao, G., Zhang, Q., Xiao, J., & He, Y. (2013). Pyramiding and evaluation of three dominant brown planthopper resistance genes in the elite indica rice 9311 and its hybrids. *Pest Management Science*, 69, 802–808. Available from <https://doi.org/10.1002/ps.3437>.
- Hu, J., Li, X., Wu, C., Yang, C., Hua, H., Gao, G., ... He, Y. (2012). Pyramiding and evaluation of the brown planthopper resistance genes *Bph14* and *Bph15* in hybrid rice. *Molecular Breeding*, 29, 61–69. Available from <https://doi.org/10.1007/s11032-010-9526-x>.
- Hu, J., Xiao, C., Cheng, M., Gao, G., Zhang, Q., & He, Y. (2015). Fine mapping and pyramiding of brown planthopper resistance genes *QBph3* and *QBph4* in an introgression line from wild rice *O. officinalis*. *Molecular Breeding*, 35, 3. Available from <https://doi.org/10.1007/s11032-015-0228-2>.
- Hussain, S., Zhu, C., Bai, Z., Huang, J., Zhu, L., Cao, X., ... Zhang, J. (2019). iTRAQ-based protein profiling and biochemical analysis of two contrasting rice genotypes revealed their differential responses to salt stress. *International Journal of Molecular Sciences*, 20, 547. Available from <https://doi.org/10.3390/ijms20030547>.
- International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*, 436, 793–800. Available from <https://doi.org/10.1038/nature03895>.
- Ishii, T., Brar, D. S., Multani, D. S., & Khush, G. S. (1994). Molecular tagging of genes for brown planthopper resistance and earliness introgressed from *Oryza australiensis* into cultivated rice, *O. sativa*. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 37, 217–221. Available from <https://doi.org/10.1139/g94-030>.
- Jairin, J., Teangdeerith, S., Leelagud, P., Phengrat, K., Vanavichit, A., & Toojinda, T. (2007). Detection of brown planthopper resistance genes from different rice mapping populations in the same genomic location. *ScienceAsia*, 33, 347–352. Available from <https://doi.org/10.2306/scienceasia1513-1874.2007.33.347>.

- Jannoey, P., Channei, D., Kotcharek, J., Pongprasert, W., & Nomura, M. (2017). Expression analysis of genes related to rice resistance against brown planthopper, *Nilaparvata lugens*. *Rice Science*, *24*, 163–172. Available from <https://doi.org/10.1016/j.rsci.2016.10.001>.
- Jena, K. K., Jeung, J. U., Lee, J. H., Choi, H. C., & Brar, D. S. (2006). High-resolution mapping of a new brown planthopper (BPH) resistance gene, *Bph18(t)*, and marker-assisted selection for BPH resistance in rice (*Oryza sativa* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *112*, 288–297. Available from <https://doi.org/10.1007/s00122-005-0127-8>.
- Ji, R., Yu, H., Fu, Q., Chen, H., Ye, W., Li, S., & Lou, Y. (2013). Comparative transcriptome analysis of salivary glands of two populations of rice brown planthopper, *Nilaparvata lugens*, that differ in virulence. *PLoS One*, *8*(11), e79612. Available from <https://doi.org/10.1371/journal.pone.0079612>.
- Ji, H., Kim, S.-R., Kim, Y.-H., Suh, J.-P., Park, H.-M., Sreenivasulu, N., ... Jena, K. K. (2016a). Erratum: Map-based cloning and characterization of the *BPH18* gene from wild rice conferring resistance to brown planthopper (BPH) insect pest. *Scientific Reports*, *6*, 36688. Available from <https://doi.org/10.1038/srep36688>.
- Ji, H., Kim, S.-R., Kim, Y.-H., Suh, J.-P., Park, H.-M., Sreenivasulu, N., ... Jena, K. K. (2016b). Map-based cloning and characterization of the *BPH18* gene from wild rice conferring resistance to brown planthopper (BPH) insect pest. *Scientific Reports*, *6*, 34376. Available from <https://doi.org/10.1038/srep34376>.
- Jing, S., Zhao, Y., Du, B., Chen, R., Zhu, L., & He, G. (2017). Genomics of interaction between the brown planthopper and rice. *Current Opinion in Insect Science*, *19*, 82–87. Available from <https://doi.org/10.1016/j.cois.2017.03.005>.
- Kabis, A., & Khush, G. S. (1988). Genetic analysis of resistance to brown planthopper in rice (*Oryza sativa* L.). *Plant Breeding*, *100*, 54–58. Available from <https://doi.org/10.1111/j.1439-0523.1988.tb00216.x>.
- Kang, K., Yue, L., Xia, X., Liu, K., & Zhang, W. (2019). Comparative metabolomics analysis of different resistant rice varieties in response to the brown planthopper *Nilaparvata lugens* Hemiptera: Delphacidae. *Metabolomics: Official Journal of the Metabolomic Society*, *15*, 62. Available from <https://doi.org/10.1007/s11306-019-1523-4>.
- Khush, G. S., Karim, A. N. M. R., & Angeles, E. R. (1985). Genetics of resistance of rice cultivar ARC10550 to Bangladesh brown planthopper tele-type. *Journal of Genetics*, *64*, 121–125. Available from <https://doi.org/10.1007/BF02931140>.
- Kumar, K., Kaur, P., Kishore, A., Vikal, Y., Singh, K., & Neelam, K. (2020). Recent advances in genomics-assisted breeding of brown planthopper (*Nilaparvata lugens*) resistance in rice (*Oryza sativa*). *Plant Breeding*, *139*, 1052–1066. Available from <https://doi.org/10.1111/pbr.12851>.
- Kumar, K., Sarao, P. S., Bhatia, D., Neelam, K., Kaur, A., Mangat, G. S., ... Singh, K. (2018). High-resolution genetic mapping of a novel brown planthopper resistance locus, *Bph34* in *Oryza sativa* L. X *Oryza nivara* (Sharma & Shastry) derived interspecific F2 population. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *131*, 1163–1171. Available from <https://doi.org/10.1007/s00122-018-3069-7>.
- Lakshminarayana, A., & Khush, G. S. (1977). New genes for resistance to the brown planthopper in rice 1. *Crop Science*, *17*, 96–100. Available from <https://doi.org/10.2135/cropsci1977.0011183X001700010028x>.
- Lang, N., & Buu, B. (2003). Genetic and physical maps of gene *Bph-10* controlling brown plant hopper resistance in rice (*Oryza sativa* L.). *Omonrice*, *11*, 35–41.
- Li, C., Luo, C., Zhou, Z., Wang, R., Ling, F., Xiao, L., ... Chen, H. (2017). Gene expression and plant hormone levels in two contrasting rice genotypes responding to brown planthopper infestation. *BMC Plant Biology*, *17*, 57. Available from <https://doi.org/10.1186/s12870-017-1005-7>.
- Li, Z., Xue, Y., Zhou, H., Li, Y., Usman, B., Jiao, X., ... Qiu, Y. (2019). High-resolution mapping and breeding application of a novel brown planthopper resistance gene derived from wild rice (*Oryza rufipogon* Griff.). *Rice*, *12*, 41. Available from <https://doi.org/10.1186/s12284-019-0289-7>.
- Liu, C., Hao, F., Hu, J., Zhang, W., Wan, L., Zhu, L., ... He, G. (2010). Revealing different systems responses to brown planthopper infestation for pest susceptible and resistant rice plants with the combined metabolomic and gene-expression analysis. *Journal of Proteome Research*, *9*, 6774–6785. Available from <https://doi.org/10.1021/pr100970q>.
- Liu, Y., Chen, L., Liu, Y., Dai, H., He, J., Kang, H., ... Wan, J. (2016). Marker assisted pyramiding of two brown planthopper resistance genes, *Bph3* and *Bph27(t)*, into elite rice cultivars. *Rice*, *9*, 27. Available from <https://doi.org/10.1186/s12284-016-0096-3>.
- Liu, Y., Wu, H., Chen, H., Liu, Y., Yanling, He, J., ... Wan, J. (2015). A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice. *Nature Biotechnology*, *33*, 301–305. Available from <https://doi.org/10.1038/nbt.3069>.
- Lv, W., Du, B., Shangguan, X., Zhao, Y., Pan, Y., Zhu, L., ... He, G. (2014). BAC and RNA sequencing reveal the brown planthopper resistance gene *BPH15* in a recombination cold spot that mediates a unique defense mechanism. *BMC Genomics*, *15*, 674. Available from <https://doi.org/10.1186/1471-2164-15-674>.
- Mohanty, S. K., Panda, R. S., Mohapatra, S. L., Nanda, A., Behera, L., Jena, M., ... Mohapatra, T. (2017). Identification of novel quantitative trait loci associated with brown planthopper resistance in the rice landrace Salkathi. *Euphytica*, *213*, 38. Available from <https://doi.org/10.1007/s10681-017-1835-2>.
- Murai, H., Hashimoto, Z., Sharma, P. N., Shimizu, T., Murata, K., Takumi, S., Mori, N., Kawasaki, S., & Nakamura, C. (2001). Construction of a high-resolution linkage map of a rice brown planthopper (*Nilaparvata lugens* Stål) resistance gene *bph2*. *Theoretical and Applied Genetics*, *103*, 526–532. Available from <https://doi.org/10.1007/s001220100598>.
- Myint, K. K. M., Fujita, D., Matsumura, M., Sonoda, T., Yoshimura, A., & Yasui, H. (2012). Mapping and pyramiding of two major genes for resistance to the brown planthopper (*Nilaparvata lugens* [Stål]) in the rice cultivar ADR52. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *124*, 495–504. Available from <https://doi.org/10.1007/s00122-011-1723-4>.
- Nanda, S., Mishra, R., & Joshi, R. K. (2021). Molecular basis of insect resistance in plants: current updates and future prospects. *Research Journal of Biotechnology*, *16*, 194–205.
- Nanda, S., Wan, P.-J., Yuan, S.-Y., Lai, F.-X., Wang, W.-X., & Fu, Q. (2018). Differential responses of *OsMPKs* in IR56 rice to two BPH populations of different virulence levels. *International Journal of Molecular Sciences*, *19*, 4030. Available from <https://doi.org/10.3390/ijms19124030>.

- Nanda, S., Yuan, S.-Y., Lai, F.-X., Wang, W.-X., Fu, Q., & Wan, P.-J. (2020). Identification and analysis of miRNAs in IR56 rice in response to BPH infestations of different virulence levels. *Scientific Reports*, *10*, 19093. Available from <https://doi.org/10.1038/s41598-020-76198-9>.
- Nemoto, H., Ikeda, R., & Kaneda, C. (1989). New genes for resistance to brown planthopper, *Nilaparvata lugens* Stal, in rice. *Ikushugaku zasshi*, *39*, 23–28. Available from <https://doi.org/10.1270/jsbbs1951.39.23>.
- Peng, L., Zhao, Y., Wang, H., Zhang, J., Song, C., Shanguan, X., ... He, G. (2016). Comparative metabolomics of the interaction between rice and the brown planthopper. *Metabolomics: Official Journal of the Metabolomic Society*, *12*. Available from <https://doi.org/10.1007/s11306-016-1077-7>.
- Prahalada, G. D., Shivakumar, N., Lohithaswa, H. C., Sidde Gowda, D. K., Ramkumar, G., Kim, S.-R., ... Jena, K. K. (2017). Identification and fine mapping of a new gene, *BPH31* conferring resistance to brown planthopper biotype 4 of India to improve rice, *Oryza sativa* L. *Rice*, *10*, 41. Available from <https://doi.org/10.1186/s12284-017-0178-x>.
- Qiu, Y., Guo, J., Jing, S., Zhu, L., & He, G. (2012). Development and characterization of japonica rice lines carrying the brown planthopper-resistance genes *BPH12* and *BPH6*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *124*, 485–494. Available from <https://doi.org/10.1007/s00122-011-1722-5>.
- Que, S., Li, K., Chen, M., Wang, Y., Yang, Q., Zhang, W., ... He, H. (2012). PhosphoRice: A meta-predictor of rice-specific phosphorylation sites. *Plant Methods*, *8*, 5. Available from <https://doi.org/10.1186/1746-4811-8-5>.
- Rahman, M. L., Jiang, W., Chu, S. H., Qiao, Y., Ham, T.-H., Woo, M.-O., ... Koh, H.-J. (2009). High-resolution mapping of two rice brown planthopper resistance genes, *Bph20(t)* and *Bph21(t)*, originating from *Oryza minuta*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *119*, 1237–1246. Available from <https://doi.org/10.1007/s00122-009-1125-z>.
- Ram, T., Deen, R., Gautam, S. K., Ramesh, K., Rao, Y. K., & Brar, D. S. (2010). Identification of new genes for brown planthopper resistance in rice introgressed from *O. glaberrima* and *O. minuta*. *Rice Genetics Newsletter*, *25*, 67–69.
- Ren, J., Gao, F., Wu, X., Lu, X., Zeng, L., Lv, J., ... Ren, G. (2016). *Bph32*, a novel gene encoding an unknown SCR domain-containing protein, confers resistance against the brown planthopper in rice. *Scientific Reports*, *6*, 37645. Available from <https://doi.org/10.1038/srep37645>.
- Renganayaki, K., Fritz, A. K., Sadasivam, S., Pammi, S., Harrington, S. E., McCouch, S. R., ... Reddy, A. S. (2002). Mapping and progress toward map-based cloning of brown planthopper biotype-4 resistance gene introgressed from *Oryza officinalis* into cultivated rice, *O. sativa*. *Crop Science*, *42*, 2112–2117. Available from <https://doi.org/10.2135/cropsci2002.2112>.
- Sangha, J., Chen, Y., Kaur, J., Khan, W., Abduljaleel, Z., Alanazi, M., ... Leung, H. (2013). Proteome analysis of rice (*Oryza sativa* L.) mutants reveals differentially induced proteins during brown planthopper (*Nilaparvata lugens*) infestation. *International Journal of Molecular Sciences*, *14*, 3921–3945. Available from <https://doi.org/10.3390/ijms14023921>.
- Sidhu, G. S., & Khush, G. S. (1978). Genetic analysis of brown planthopper resistance in twenty varieties of rice, *Oryza sativa* L. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *53*, 199–203. Available from <https://doi.org/10.1007/BF00277368>.
- Su, C., Zhai, H., Wang, C., Sun, L., & Wan, J. (2006). SSR Mapping of brown planthopper resistance gene *Bph9* in Kaharamana, an Indica rice (*Oryza sativa* L.). *Acta Genetica Sinica*, *33*(3), 262–268. Available from [https://doi.org/10.1016/S0379-4172\(06\)60049-8](https://doi.org/10.1016/S0379-4172(06)60049-8).
- Sun, L., Su, C., Wang, C., Zhai, H., & Wan, J. (2005). Mapping of a major resistance gene to the brown planthopper in the rice cultivar Rathu Heenati. *Breeding Science*, *55*, 391–396. Available from <https://doi.org/10.1270/jsbbs.55.391>.
- Tamura, Y., Hattori, M., Yoshioka, H., Yoshioka, M., Takahashi, A., Wu, J., ... Yasui, H. (2015). Map-based cloning and characterization of a brown planthopper resistance gene *BPH26* from *Oryza sativa* L. ssp. indica cultivar ADR52. *Scientific Reports*, *4*, 5872. Available from <https://doi.org/10.1038/srep05872>.
- Tan, J., Wu, Y., Guo, J., Li, H., Zhu, L., Chen, R., ... Du, B. (2020). A combined microRNA and transcriptome analyses illuminates the resistance response of rice against brown planthopper. *BMC Genomics*, *21*, 144. Available from <https://doi.org/10.1186/s12864-020-6556-6>.
- Uawisetwathana, U., Chevallier, O. P., Xu, Y., Kamolsukyeunyoung, W., Nookaew, I., Somboon, T., ... Karoonuthaisiri, N. (2019). Global metabolite profiles of rice brown planthopper-resistant traits reveal potential secondary metabolites for both constitutive and inducible defenses. *Metabolomics: Official Journal of the Metabolomic Society*, *15*, 151. Available from <https://doi.org/10.1007/s11306-019-1616-0>.
- Wan, P. J., Yuan, S. Y., Wang, W. X., Chen, X., Lai, F. X., & Fu, Q. (2016). A genome-wide identification and analysis of the basic helix-loop-helix transcription factors in brown planthopper, *Nilaparvata lugens*. *Genes*, *7*, 100. Available from <https://doi.org/10.3390/genes7110100>.
- Wang, H., Shi, S., Guo, Q., Nie, L., Du, B., Chen, R., ... He, G. (2018). High-resolution mapping of a gene conferring strong antibiosis to brown planthopper and developing resistant near-isogenic lines in 9311 background. *Molecular Breeding*, *38*, 107. Available from <https://doi.org/10.1007/s11032-018-0859-1>.
- Wang, W. X., Lai, F. X., Wan, P. J., Fu, Q., & Zhu, T. H. (2019). Molecular characterization of Ca²⁺/calmodulin-dependent protein kinase II isoforms in three rice planthoppers—*Nilaparvata lugens*, *Laodelphax striatellus*, and *Sogatella furcifera*. *International Journal of Molecular Sciences*, *20*, 3014. Available from <https://doi.org/10.3390/ijms20123014>.
- Wang, Y., Cao, L., Zhang, Y., Cao, C., Liu, F., Huang, F., ... Lou, X. (2015). Map-based cloning and characterization of *BPH29*, a B3 domain-containing recessive gene conferring brown planthopper resistance in rice. *Journal of Experimental Botany*, *66*, 6035–6045. Available from <https://doi.org/10.1093/jxb/erv318>.
- Wei, Z., Hu, W., Lin, Q., Cheng, X., Tong, M., Zhu, L., ... He, G. (2009). Understanding rice plant resistance to the brown planthopper (*Nilaparvata lugens*): A proteomic approach. *Proteomics*, *9*, 2798–2808. Available from <https://doi.org/10.1002/pmic.200800840>.
- Wu, H., Liu., Yuqiang, He, J., Liu, Yanling, Jiang, L., Liu, L., ... Wan, J. (2014). Fine mapping of brown planthopper (*Nilaparvata lugens* Stål) resistance gene *Bph28(t)* in rice (*Oryza sativa* L.). *Molecular Breeding*, *33*, 909–918. Available from <https://doi.org/10.1007/s11032-013-0005-z>.
- Wu, Y., Lv, W., Hu, L., Rao, W., Zeng, Y., Zhu, L., ... He, G. (2017). Identification and analysis of brown planthopper-responsive microRNAs in resistant and susceptible rice plants. *Scientific Reports*, *7*, 8712. Available from <https://doi.org/10.1038/s41598-017-09143-y>.

- Yang, H., You, A., Yang, Z., Zhang, F., He, R., Zhu, L., & He, G. (2004). High-resolution genetic mapping at the *Bph15* locus for brown planthopper resistance in rice (*Oryza sativa* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *110*, 182–191. Available from <https://doi.org/10.1007/s00122-004-1844-0>.
- Yang, M., Cheng, L., Yan, L., Shu, W., Wang, X., & Qiu, Y. (2019). Mapping and characterization of a quantitative trait locus resistance to the brown planthopper in the rice variety IR64. *Hereditas*, *156*, 22. Available from <https://doi.org/10.1186/s41065-019-0098-4>.
- Yang, Z., Zhang, F., Zhu, L., & He, G. (2006). Identification of differentially expressed genes in brown planthopper *Nilaparvata lugens* (Hemiptera: Delphacidae) responding to host plant resistance. *Bulletin of Entomological Research*, *96*, 53–59. Available from <https://doi.org/10.1079/BER2005400>.
- Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., . . . Yang, H. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science (New York, N.Y.)*, *296*, 79–92. Available from <https://doi.org/10.1126/science.1068037>.
- Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., . . . Yang, H. (2005). The genomes of *Oryza sativa*: A history of duplications. *PLoS Biology*, *3*, e38. Available from <https://doi.org/10.1371/journal.pbio.0030038>.
- Zhang, F., Zhu, L., & He, G. (2004). Differential gene expression in response to brown planthopper feeding in rice. *Journal of Plant Physiology*, *161*, 53–62. Available from <https://doi.org/10.1078/0176-1617-01179>.
- Zhang, X., Yin, F., Xiao, S., Jiang, C., Yu, T., Chen, L., . . . Li, W. (2019). Proteomic analysis of the rice (*Oryza officinalis*) provides clues on molecular tagging of proteins for brown planthopper resistance. *BMC Plant Biology*, *19*, 30. Available from <https://doi.org/10.1186/s12870-018-1622-9>.
- Zhang, Y., Qin, G., Ma, Q., Wei, M., Yang, X., Ma, Z., . . . Li, R. (2020). Identification of major locus Bph35 resistance to brown planthopper in rice. *Rice Science*, *27*(3), 237–245. Available from <https://doi.org/10.1016/j.rsci.2020.04.006>.
- Zhao, Y., Huang, J., Wang, Z., Jing, S., Wang, Y., Ouyang, Y., . . . He, G. (2016). Allelic diversity in an NLR gene *BPH9* enables rice to combat planthopper variation. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 12850–12855. Available from <https://doi.org/10.1073/pnas.1614862113>.
- Zhou, C., Zhang, G., Noman, W., Li, C., Zhou, L., & Lou. (2019). OsMKK3, a stress-responsive protein kinase, positively regulates rice resistance to *Nilaparvata lugens* via phytohormone dynamics. *International Journal of Molecular Sciences*, *20*, 3023. Available from <https://doi.org/10.3390/ijms20123023>.

This page intentionally left blank

Contemporary genomic approaches in modern agriculture for improving tomato varieties

Nikolay Manchev Petrov¹, Mariya Ivanova Stoyanova², Rajarshi Kumar Gaur³, Milena Georgieva Bozhilova-Sakova⁴ and Ivona Vassileva Dimitrova⁵

¹Department of Natural Sciences, New Bulgarian University, Sofia, Bulgaria, ²Department of Plant Protection, Institute of Soil Science, Agrotechnologies and Plant Protection (ISSAPP) “Nikola Pushkarov”, Sofia, Bulgaria, ³Department of Biotechnology, Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India, ⁴Department “Genetics, Breeding, Selection, Reproduction and Biotechnologies of Farm Animals”, Institute of Animal Science, Kostinbrod, Bulgaria, ⁵Department of Plant Protection, Agronomy Faculty, University of Forestry, Sofia, Bulgaria

17.1 Importance and origin of tomatoes

Lycopersicon esculentum Mill, the cultivated tomato, is widely grown around the world. According to the latest data from FAOSTAT, the worldwide production of tomatoes amounts to 182 million tons in 2018, and for the period 1988–2018 it has increased 2.7 times (FAOSTAT, 2018; <http://faostat.fao.org>). Tomato is the second most-consumed vegetable after potato (FAOSTAT, 2018; <http://faostat.fao.org>). However, unlike potatoes, tomatoes can be consumed both fresh and processed in a variety of forms. Major tomato-producing countries are China, the United States, India, and Turkey (<http://faostat.fao.org>).

Tomato fruits contain 95% water, 3%–4% carbohydrates, 1% protein, and 0.2% lipids. They are low in calories (only 16–18 kcal/100 g) but are rich in vitamins and trace elements, which makes them suitable for different types of diets. Tomatoes contain calcium, iron, magnesium, phosphorus, potassium, sodium, zinc, copper, manganese and selenium, vitamin A (retinol), vitamin E (α -tocopherol), vitamins from the B-group (thiamine, riboflavin, niacin, choline, pantothenic acid, pyridoxine), and vitamin K (phylloquinone). Vitamin C is in high content ($\sim 1/3$ of RDA/100 g). Tomatoes also contain all nine essential amino acids, although in modest amounts. Tomatoes with different fruit colors have minimal differences in composition except for the ratio of carotenoids. Some varieties have been selected with an increased content of vitamin C and carotenoids, which are one of the most important indicators of the quality of the variety (Framar, 2016; USDA, 2016). Most interesting from biochemical and medical point of view are the compounds from the groups of peptides, alkaloids, and carotenoids.

Solanum lycopersicum contains the alkaloids α -tomatin and dehydrotomatin in different proportions in all plant parts except the ripe fruits. α -Tomatin inhibits the growth of fungi by binding to 3β -hydroxy sterols in their membranes and both the alkaloids are toxic to animal cell lines (Kozukue, Han, Lee, & Friedman, 2004). Although tomatin is considered toxic, its concentration in the leaf mass and green fruits is too low to be dangerous. Moreover, it was found that it binds to cholesterol in the digestive tract in mice and is excreted from the body (Barceloux, 2009; McGee, 2009). Tomato plants contain also solanine—another toxic alkaloid but its concentration in green fruits is even lower compared to potatoes (Barceloux, 2009; McGee, 2009). Esculeoside A (contained in rose fruits), esculeoside B (contained in red fruits), and lycoperoside (contained in fruits and leaves) are other alkaloids of steroidal nature in tomatoes. While ripening, the content of α -tomatin, dehydrotomatin, and solanine in the fruits decreases, and that of esculeoside increases (Eltayeb & Roddick, 1984; Katsumata et al., 2011; Kozukue et al., 2004; Moco et al., 2007). Esculeoside A can be metabolized to various steroid hormones used in the human body and thus have a beneficial effect on human health. Cherry tomatoes are particularly rich in esculeoside. Lycoperoside is higher in immature fruits but is also present in ripe fruits (Manabe et al., 2011; Moco et al., 2007). Tomato fruits contain various phenolic compounds

(*p*-hydroxybenzoic acid, naringenin, kaempferol, quercetin, myricetin, astragalín, and some of their derivatives) which are very effective neutralizers of peroxide radicals due to their chemical structure. In addition, they can chelate iron ions that catalyze lipid oxidation (Martinez-Valverde, Periago, Provan, & Chesson, 2002; Nice, 2013). Phenolic compounds are natural preservatives due to their antibacterial or/and antifungal effect (Eklund, 1985; Nice, 2013). Naringenin derivative—naringenin chalcone inhibits the release of histamine thus having an antiallergic effect (Yamamoto et al., 2004; Nice, 2013). Quercetin is contained in good amounts in ripe fruits and has a complex action—strong antibacterial and antifungal effect, combined with antiinflammatory effect in humans. In addition, this substance is an antioxidant and reduces the risk of cancer (Nice, 2013; Omodamiro & Amechi, 2013). Astragalín has both antimicrobial activity, antiallergic and antiinflammatory, and effect (Nice, 2013).

Despite the rich content of vitamins, minerals, and phenolic acids, the most valued in tomato fruit are pigments. Pigments are found in different colors: red (lycopene), orange (β -carotene), yellow-orange (zeaxanthin), yellow (lutein), green (chlorophyll), blue-violet (anthocyanins) (Kong et al., 2010). Anthocyanins are found mainly in purple varieties of tomatoes. Their color actually varies from red to purple depending on the concentration and pH. They are usually in the form of glycosides. They are known for their antioxidant properties but also have other effects: antiinflammatory, immunomodulatory, and immunostimulating. Plants are thought to synthesize anthocyanins as protective molecules against environmental stressors—ultraviolet light, low temperatures, drought (Ghosh & Konishi, 2007; Riaz, Zia-Ul-Haq, & Saad, 2016; Webb, 2014). Lutein and zeaxanthin are two similar carotenoid isomers that are not involved in the synthesis of vitamin A. Both the pigments accumulate in the macula of the eye, with the main role to absorb ultraviolet rays, protecting the inner layers of the eye. It has been found that diets rich in lutein and zeaxanthin prevent the development of macular degeneration, eye cataracts, and diabetic retinopathy and improve night vision (Heiting, 2014; Koushan, Rusovici, Li, Ferguson, & Chalam, 2013; Stankovic, 2004). Tomatoes rank third in the content of the red pigment lycopene, with red tomatoes accumulating six times more than red peppers (Cooper & Nicola, 2014; HealthAliciousNess, 2016; Kong et al., 2010). Both red and orange tomatoes contain lycopene and β -carotene, and the final color of the fruit depends on the ratio between these two main pigments (Martinez-Valverde et al., 2002). Both lycopene and β -carotene are composed of hydrogen and 40 carbon atoms. But unlike β -carotene, lycopene has a linear structure and two additional C–C bonds that increase light absorption and turn orange to red. There are several common isomers of lycopene, and the possible isomers are 72 (Cooper & Nicola, 2014; Martinez-Valverde et al., 2002). However, in tomato fruits, about 95% of lycopene is in trans form, while with higher bioavailability is cis form. The absorption of lycopene in the human body depends on various factors, temperature treatment, light, and concomitant consumption of sunflower oil being some of them. Studies have suggested that the acidic environment in the stomach partially catalyzes the isomerization of lycopene from trans to cis form. Sun-dried tomatoes provide the highest bioavailability of lycopene compared to raw and canned tomatoes (Kong et al., 2010). Lycopene has various beneficial effects on the human body. It is the most effective carotenoid, neutralizing oxygen- and peroxide-free radicals, mainly due to a large number of double bonds in its molecule (Kong et al., 2010; Martinez-Valverde et al., 2002). Lycopene is two times more effective than β -carotene in neutralizing free radicals and 10 times more effective than α -tocopherol (vitamin E) (Kong et al., 2010). The main mechanism by which antioxidants (such as phenols, vitamin E, and flavonoids) inhibit lipid autooxidation is the elimination of the peroxide radical by delivering a hydrogen atom and the formation of a lipid peroxide and a resonance-stabilized antioxidant radical (Sies & Stahl, 1995). As a carotenoid, lycopene can eliminate free radicals through other mechanisms. The combinations of lycopene and other antioxidants contained in tomato fruit have synergistic effect in exhibiting antioxidant properties (Kong et al., 2010). Various studies have shown that diets rich in lycopene reduce the likelihood of developing various types of cancer, cardiovascular disease, atherosclerosis, and neurodegenerative diseases. A chemoprotective effect, as well as antibacterial and antifungal properties have also been established (Kong et al., 2010; Krishna, Bhaumik, & Kumar, 2013; Martinez-Valverde et al., 2002; Omodamiro & Amechi, 2013). The content of lycopene in tomatoes varies greatly depending on several factors. Varieties containing the so-called crimson genes accumulate higher pigment content. Temperatures also have an effect, with optimal temperatures between 16°C and 20°C, and temperatures above 30°C inhibiting lycopene synthesis (Martinez-Valverde et al., 2002).

S. lycopersicum is very adaptive and is grown in almost every part of the world from the tropics to within a few degrees of the Arctic Circle in the fields or in greenhouses (<http://faostat.fao.org>). Among all *Lycopersicon* species, only *L. esculentum* has become a domesticated crop (Peralta & Spooner, 2007; Rick, 1978). The wild species *Solanum pimpinellifolium* (with fruit diameter \sim 1 cm) is also casually planted for consumption in Peru. Crossing the cultivated tomato with this species in the selection aims to improve the color and qualities of the fruit, as well as to inherit disease resistance. *Solanum cheesmaniae* is endemic to the Galapagos Islands with yellow, orange to purple fruits. It is characterized by salt tolerance and resistance to viral diseases. A closely related and also endemic to the region species with

orange fruits is *Solanum galapagense*, which grows mainly along the coast, near seawater, and enters the territories only up to 50 m above sea level. Nine other wild species with green fruits have been described. They are adapted to different habitats—dry soils, rocky or wet places, and their preferences for altitude also vary widely. Highly adaptable and drought-resistant is *Solanum chilense* with purple fruits, which grows from 0 to 3250 m above sea level (Peralta & Spooner, 2007) and is also used for improving the resistance and drought-tolerance in *S. lycopersicum*. All of these species grow naturally in Peru, which is the reason why tomato species were traditionally thought to origin from this country (Juvik, Berlinger, Ben-David, & Rudich, 1982; Rick, 1976a, 1976b, 1978).

The first systematic study was done by DeCandolle in 1886, who combined data from botany, history, and philology. DeCandolle believed that the homeland of tomatoes is the west coast of the continent and the Galapagos Islands, and the ancestor of the cultivated tomato is a wild tomato with very small fruits, reaching only 2.5 cm in diameter. The ancient inhabitants of Peru cultivated it before the discovery of America by Christopher Columbus. In 1623 Bauhin described the tomato as “mala peruviana” and “pomi del Peru” (Peruvian apple), which suggests the transfer of the plant from Peru to Europe. Peruvian origins were supported by other authors in the first half of the 20th century (Peralta & Spooner, 2007). The first record of tomatoes in Europe is credited to descriptions published in 1554 by Italian herbalist Mattheolus who made a description and an illustration of the plant and called it “pomi d’oro” or in Latin—“mala aurea” (golden apple). The plant was undoubtedly a tomato, but without reference to a geographical origin (Jenkins, 1948; Rick, 1978). These and equivalent names persisted well into the 19th century. Seventeen years later, Anguillara first used the name “pomi del Peru” together with “pomi d’oro,” but it is not clear whether he called the same plant by those names. According to Jenkins, “pomi del Peru” actually refers to other plants of the *Solanaceae* family, such as *Datura stramonium* (tatul) (Jenkins, 1948). The author also points out that there are data on the early cultivation of tomatoes in Mexico in the pre-Columbian era, while such data for South America are lacking (Jenkins, 1948). At the same time, biodiversity in native Mexican varieties was claimed to be significantly higher than in Peruvian varieties (Rick, 1976a, 1976b, 1991). Vavilov and Jenkins suggested that it was possible that the wild relative of the cultivated tomato was transported to Mexico, where it was cultivated in pre-Columbian times before, later transported to Europe (Jenkins, 1948; Peralta & Spooner, 2007). However, greater varietal diversity in Mexico has not been confirmed by later comparative studies (Villand et al., 1998). At the same time, in 1975, Rick and Fobes found that tomatoes from Europe and North America had the same isozymes as those in Mexico and Central America (Peralta & Spooner, 2007). Comparative genetic studies based on RAPD (random amplified polymorphic DNA) and restriction fragment length polymorphisms (RFLP) (restriction fragment length polymorphism) of old native varieties from Mexico, Central America, and Peru, neither support nor deny the hypothesis of Mexican origin (Rick & Holle, 1990; Villand et al., 1998; Williams & St. Clair, 1993). The first contact of Europeans with Mexico was in 1519 during the conquest of Mexico City, and with Peru—12 years later during the conquest of Peru. Botanists at that time were mainly interested in the medicinal and culinary applications of plants. Due to its close relationship with poisonous plants of the *Solanaceae* family, such as mandrake (*Mandragora* sp.) and dog grape (*Solanum nigrum*), the plant was originally considered poisonous and was grown for decoration (Peralta & Spooner, 2007). Consumption of tomatoes began first in the south, while in the northeastern regions it happened in the 19th century (Peralta & Spooner, 2007). It is arguably accepted that the wild cherry (*L. esculentum* var. *cerasiforme*, with fruit diameter of ~1.5–3 cm) is the immediate progenitor of the cultivated tomato though *Lycopersicon pimpinellifolium* is also a likely candidate (Jenkins, 1948; Rick, 1976a, 1976b).

S. lycopersicum is characterized by many varieties and cultivars. Plants can be of limited growth (determinant varieties) or of unlimited growth (indeterminate varieties grown with stakes). Fruits come in a wide range of shapes, sizes, and colors. Some may be globe, round, flattened, oval, oblong, heart-shaped, lemon-shaped, pear-shaped, even pepper-shaped, and some varieties have an interesting decorative shape, such as the varieties “Heart of Albenga” and “Rose,” which have vertical ribs (Varietal Seeds and Plant Protection Ltd., 2016). The size of the fruits also varies greatly between different varieties—from 7 to 8 g for some varieties of cherry to 2 kg for the variety Gigant (Bulgarian Farmer, 2016; Varietal Seeds and Plant Protection Ltd., 2016). Their colors may be red, raspberry-red, pink, red-orange, orange, golden, yellow, purple, green, striped, or white/ivory (Bulgarian Farmer, 2016; Varietal Seeds and Plant Protection Ltd., 2016). There are more varieties of tomato sold worldwide than any other vegetable.

17.2 Organization of tomato genome and genetic variation of tomato cultivars

Along with the modern tomato (*S. lycopersicum* L. var. *esculentum*) and its wild form [*S. lycopersicum* L. var. *Cerasiforme* (Dun.)], there are eight other related wild species—*S. pimpinellifolium*, *S. cheesmanii*, *S. chmielewskii*, *S. chilense*, *S. parviflorum*, *S. peruvianum*, *S. hirsutum*, *S. pennellii*, which occur as native forms in Peru, Western South America, and the Caribbean (Rick, 1976a, 1976b, 1979). European tomato varieties were introduced by Spanish

researchers, and the population of *S. lycopersicum* shows about 5% genetic variability (Miller & Tanksley, 1990; Rick & Fobes, 1975). The natural habitats of tomatoes as diploid species with similar chromosomes vary widely, from dry to wet and from flat to high mountain geographical places (Rick & Butler, 1956; Warnock, 1988). Many of these different tomato species create a variety of qualities that can be used to improve the desired qualities of the cultivated tomato. Those of them that have problems with sexual hybridization can cross by embryo rescue techniques or by pollen mixture (Rick, 1973, 1976a, 1979; Scott, Jones, & Somodi, 1995; Scott, Olson, et al., 1995). Resistance to tobacco mosaic virus (TMV) and nematodes has been achieved in some of the crosses. Thus, these wild tomato species represent a valuable genetic source for improving the qualities of the cultivated tomato (Bretó, Asins, & Carbonell, 1993; Miller & Tanksley, 1990). During the long-term selection, many wild species with desired characteristics have been selected, such as better quality fruits, abiotic stress resistance, diseases, and pest resistance. The improvement and development of molecular biology techniques and the discovery of new molecular markers would facilitate the selection process, with thousands of tomato species and varieties being stored in genetic banks in many countries around the world (Rossi et al., 1998).

The tomato genome is represented by approximately 35,000 genes, with an approximate gene density of 6.7 kb/gene (Khush & Rick, 1968; Peterson, Pearson, & Stack, 1998; van der Hoeven, Ronning, Giovannoni, Martin, & Tanksley, 2002; Wang, van der Hoeven, Nielsen, Mueller, & Tanksley, 2005). The nuclear genome of the tomato is 950 Mbp of DNA located on 12 chromosomes, with only about 30% encoding, 60% noncoding information, and 10% transposons. A large proportion of tomato euchromatin is in the methylated state in the intergenic regions (Wang et al., 2005). About two-thirds of the chromosomal DNA of a tomato are represented by heterochromatic regions that do not contain genes, with the genes located in the distal euchromatin regions (Khush & Rick, 1968; Wang, van der Hoeven, & Nielsen, 2006).

Genetic libraries taken from wild species are used to improve the quality of cultivated tomato varieties (Zamir, 2001). Such DNA libraries are introgression lines (IL) that are derived from interspecific crosses involving much of the wild genome. By using reverse genetics and screening for quantitative trait loci (QTLs) from populations of IL, candidate genes of such desired trait characteristics are identified (Ballester et al., 2016). This allows finding the loci and cloning of the desired QTL and genes, genetic links to the transcriptome, metabolome, and evaluation of a selected specific set of genes, avoiding the occurrence of wild-type gene segregation and epistasis. Thus the loci of genetic areas responsible for vegetative traits such as plant viability, as well as those related to fruits—organoleptic quality, morphology, color, and secondary metabolism in the tomato fruit wig can be located (Alseekh et al., 2015; Barrantes et al., 2016; Hanson et al., 2007). To avoid the long selection process in IL and to improve the quality of the donor genome, single-nucleotide polymorphism (SNP) genotyping is created (Barrantes et al., 2014).

Nowadays, we have the most diverse and numerous genetic resources obtained through sequencing, transcriptome, and metabolic data of scientists from many countries (Calafiore et al., 2016). Tomatoes as a species have a relatively small genome composed of 24 acrocentric to metacentric chromosomes with a high level of homozygosity, containing about 32,000 genes (Anderson, Covey, Larsen, Bedinger, & Stack, 2010; El-Awady, El-Tarras, & Hassan, 2012; Koo et al., 2008; The Tomato Genome Consortium, 2012). The tomato genome is represented by 77% heterochromatin and 23% euchromatin (Peterson, Stack, Price, & Johnston, 1996). Tomatoes can be grown in a variety of conditions, have a short life cycle of 70–90 days, produce many seeds and its sexual reproduction is easily controlled. It can also be asexually propagated by grafting or tissue culture, with a high regenerative capacity (Gerszberg, Hnatuszko-Konka, Kowalczyk, & Kononowicz, 2015). With the development of genetic engineering, the accumulation of many genetic and genomic resources, and cytogenetic research, tomato selection is one of the most advanced areas in agriculture, with methods available for the effective transformation of tomatoes (Calafiore et al., 2016).

One of the first plants for which a molecular map was created was the tomato (Foolad, 2007a, 2007b). These maps are used to identify gene loci and chromosome segments in different tomato species and to identify genetic changes and mutations (Karp, 2002). This makes it possible to study the chromosomal locations of the QTL genes to improve the quantity and quality of tomato production (Frary et al., 2000). Various molecular maps of tomatoes, with more than 2200 identified loci, currently exist using methods such as amplified fragment length polymorphisms (AFLP), cleaved amplified polymorphic sequences (CAPS), RFLP, RAPD, simple sequence repeats (SSR), and polymerase chain reaction (PCR) codominant markers (Shirasawa et al., 2010). With the help of population genetics and germplasm genotyping, allelic diversity was detected by genome scanning (Labate, Robertson, & Baldo, 2009).

The nucleotide sequence of the tomato genome is established during the transition from the era of sequencing by the method of multiparallel sequencing of Sanger to the era of new-generation sequencing (NGS). The tomato genome sequencing project started in 2004 as a consortium of 10 countries (Korea, China, Great Britain, India, the Netherlands, France, Japan, Spain, Italy, and the United States). Sequencing initially included a BAC-by-BAC (bacterial artificial chromosome) approach, which was successfully applied to previous model plants, and three BAC libraries of 30,800

BAC clones were built—based on the enzymes EcoRI, MboI, and HindIII (Peters et al., 2006; Sato et al., 2008). In 2009 the introduction of 3 NGS sequencing platforms began—454 (Margulies et al., 2005), SOLiD (McKernan et al., 2009), and Illumina (Harris et al., 2008). NGS has also been used successfully for transcriptional sequencing (Nagalakshmi et al., 2008).

The detection of mutations in selected genes is greatly facilitated by the use of the TILLING method (Targeting Induced Local Lesions in Genomes), which accelerates the functional genomic analysis of organisms. NGS identifies mutations in the population of *S. lycopersicum*. About 25 genes responsible for the metabolism of carotenoids and folate have been PCR-amplified and screened to find potentially useful alleles for improving tomato quality. Various software programs such as CAMBA, CRISP, GATK UNIFIED GENOTYPER, LOFREQ, SNVER, and VIPR have been used to predict mutations in the tomato genome. False-positive results are eliminated by using more than two different software programs. Screening of the 23.47 Mb tomato genome predicted 75 mutations, 64 of which were confirmed by sequencing with an average mutation density of 1/367 Kb.

17.3 Tomato breeding

Breeding new varieties of tomato started more than 200 years ago in Europe (mainly in Italy). Breeding included meeting of different needs, including fresh market and processing industries. Other important goals were disease resistance, earliness in maturity, and adaptability to different climatic conditions, for example, tolerance to adverse temperatures, resistance to rain-induced cracking while retaining and developing nutrition and taste quality (Stevens & Rick, 1986; Tigchelaar, 1986). However, yield remains aim number one for breeders. Generally, for a new variety to be successful, it has to guarantee at least the same if not exceeding yield potential, even if it possesses other improved characteristics (Warren, 1998). Since yield is an aggregate result of individual traits, breeding is usually pointed toward these traits—for example, weight of fruits, disease resistance, heat tolerance, all contribute to improved yield (Scott, 1993; Scott, Bryan, & Ramos, 1997; Scott, Miller, & Stall, 1997).

Traditional breeding is based on phenotypic selection and progeny testing, and its efficacy in improving crop productivity and fruit quality has been proven through the years (Duvick, 1986, 1996; Warren, 1998). However, the major drawback of these methods is time consumption and difficulties. Often desired traits, such as disease and pest resistance, abiotic stress tolerance, and improved fruit quality, can be acquired only from wild species. The process of gaining the necessary genes encounters a variety of problems. After interspecific hybridization, a number of undesirable genes introduced from the wild donor have to be eliminated. This includes a series of backcrosses to the cultivated parent which can occasionally lead to limiting the expression of the desired genes or even their omission. For these reasons the traditional breeding of a cultivar takes somewhere between 10 and 15 years.

In the past, breeding was based predominantly on developing open-pollinated inbred cultivars and their use for commercial production. Since the 1970s, however, F1 hybrids have become a significant part of the production. The use of hybrids in tomato allows combining of important and valuable traits and protection of breeders' work (Foolad, 2007a).

17.4 Disease resistance

The selection of disease-resistant tomato varieties is a major goal of most tomato-growing programs due to the rapid emergence of new breeds and strains of existing pathogens. The ultimate goal is to reduce the use of pesticides in tomato production by screening for disease resistance. More than 110 pathogenic species cause about 160 serious diseases in tomato cultivars, including plant pathogenic fungi, bacteria, and viruses (Tanksley & Fulton, 2007).

Disease control is crucial to prevent economic losses in fresh tomato production as well as industry (<http://faostat.fao.org>). Resistance to bacterial and fungal pathogens such as powdery mildew and bacterial wilt is a horizontal polygenic resistance without established resistance genes. Resistance resources to diseases have been found in most wild tomato species, such as *L. pimpinellifolium*, *L. peruvianum*, and *L. hirsutum*. Both vertical and horizontal resistance have been identified for some tomato fungal diseases caused by *Phytophthora infestans* and *Oidium lycopersicum* (Scott & Jones, 1986; Scott, Francis, Miller, Somodi, & Jones, 2003; Scott, Bryan, et al., 1997; Scott, Jones, et al., 1995; Scott, Miller, et al., 1997; Scott, Olson, et al., 1995; Stommel & Zhang, 1998; Yang, Sacks, Lewis, Miller, & Francis, 2005).

About 770 datasets for pathogen recognition genes are categorized according to their loci in the genome, the presence and arrangement of protein domains, and phylogenetic analysis (Andolfo et al., 2013). These are the genes for resistance to cortical root rot—Py1 located on chromosome 3, the gene for resistance to *Pseudomonas syringae*—Pto on chromosome 5, the gene for resistance to Tomato yellow leaf curl virus—Ty1 on chromosome 6 and the gene for

resistance to Tomato spotted wilt virus—Sw5 on chromosome 9 (Doganlar, Frary, Daunay, Lester, & Tanksley, 2002; Hanson et al., 2000; Martin, Williams, & Tanksley, 1991; Stevens, Lamb, & Rhoads, 1995). Some of these resistance genes have been cloned using genetic map-based methods, and many others have been localized but not cloned (Ercolano, Sanseverino, Carli, Ferriello, & Frusciante, 2012; Foolad, 2007b).

17.5 Insect resistance

Tomatoes are attacked by various pests, causing great losses such as insects, including mites, whiteflies, aphids, representatives of *Lepidoptera*, *Coleoptera*, thrips, and worms. Unlike tomato varieties developed for disease resistance, pest-resistant varieties are significantly fewer. Resistance to the main insect pests on tomatoes has been established within the isolated wild species *S. hirsutum* and *S. pennellii* (Farrar, Barbour, & Kennedy, 1994; Juvik et al., 1982; Muigai, Bassett, Schuster, & Scott, 2003; Mutschler et al., 1996; Schalk & Stoner, 1976; Tigchelaar, 1986; Weston, Johnson, Burton, & Snyder, 1989). *L. hirsutum* is a source of resistance to more than 16 species of arthropods (Farrar & Kennedy, 1991; Weston et al., 1989). Resistance to more than nine species, including greenhouse whitefly and potato aphid, has been observed in *S. pennellii* (Muigai et al., 2003). The species *S. lycopersicum* var. *cerasiforme*, *S. pimpinellifolium*, *S. cheesmanii*, *S. chmielewskii*, *S. peruvianum*, and *S. chilense* also have some resistance to insect pests (Farrar & Kennedy, 1991).

17.6 Abiotic stress tolerance

The growth and development of cultivated tomatoes are significantly sensitive to various environmental stressors such as drought, moisture, salinity, temperature, pollutants, and lack of minerals. There are few cultivated species that are resistant to various adverse environmental factors. Some wild species such as *S. chilense*, *S. peruvianum*, *S. pennellii*, *S. pimpinellifolium*, *S. hirsutum*, *S. cheesmanii*, *S. chmielewskii*, *S. rickii*, *S. juglandifolium*, *S. ochranthum*, and *S. parviflorum* are sources of such resistance to abiotic stress (Foolad, 2005; Rick, DeVerna, Chetelat, & Stevens, 1987; Rick, DeVerna, & Chetelat, 1990; Rick, 1988). The plant's response to adverse conditions is due to various physiological and anthropogenic factors that are controlled by different genes, whose expression is also influenced by environmental factors. In addition, tolerance during one stage of plant development is not related to tolerance during other stages of development (Asins, Bretó, Cambra, & Carbonell, 1993; Foolad & Chen, 1998; Foolad, Chen, & Lin, 1998; Foolad, 1999; Jones & Qualset, 1984). To ensure efficient production of tomatoes under abiotic stress, tolerance is needed at all major stages of plant development from seed germination, vegetative growth, flowering, and fruit ripening.

In the case of tomatoes, significant progress has been made in the selection of varieties tolerant to temperature influences. Highly temperature-resistant commercial tomato varieties have been successfully developed (Scott, Everett, & Bryan, 1985; Scott, Volin, Bryan, & Olson, 1986; Scott, Jones, et al., 1995; Scott, Olson, et al., 1995). QTLs responsible for salt resistance during seed germination, vegetative growth, and later stages of the life cycle have been identified. Such genes have been found in different wild tomato species and at different concentrations of saline. Such QTLs were found in crosses between *S. lycopersicum* × *S. pennellii* and *S. lycopersicum* × *S. pimpinellifolium* (Foolad & Chen, 1998; Foolad, Stoltz, Dervinis, Rodriguez, & Jones, 1997). These QTLs from different populations of the same cross show stability over generations. Also, the same QTLs contribute to tolerance at different levels of salt stress (Foolad & Jones, 1991).

Several QTLs of interspecific crossings between *S. lycopersicum* and *S. pimpinellifolium* have been established during seed germination for cold tolerance. The results of these studies show that resistance to cold during seed germination in tomatoes is controlled by more than one gene. Comparison of such QTLs in different populations of the same cross shows conservatism of detected QTLs in different interspecific tomato populations, including those derived from crosses between *S. lycopersicum* × *S. pennellii* and *S. lycopersicum* × *S. pimpinellifolium* (Foolad & Chen, 1998).

There are large variations in the shape of the fruit in the cultivated tomato, including round, ovoid, heart-shaped, elliptical, plum-shaped, elongated, and pear-shaped. Several genes responsible for fruit shape have been identified, such as pr (pyriform), o (ovate), bk (beaked tomato), n (nipple-tip tomato), f (fasciated), and lc (for locule number) (Young & MacArthur, 1947). The QTL responsible for fetal shape (called fs8.1) was located on chromosome 8 and later cloned and characterized (Grandillo, Ku, & Tanksley, 1996; Ku, Grandillo, & Tanksley, 2000). fs8.1 changes the length of the fruit during parenthesis, resulting in longer and larger ripe fruits. Thus another fetal shape QTL controls the transition from round to pear-shaped fruit (Ku, Doganlar, Chen, & Tanksley, 1999). Variation in fruit shape is controlled by several major loci due to allelic variation in these loci (Ku et al., 2000).

Fruit color is an important characteristic of the quality of fruits in tomatoes due to the growing demands of the consumer for health benefits. The presence of lycopene, the main carotenoid in tomatoes, which is responsible for the red color of the fruit, has become a desirable feature (Gerster, 1997; Mascio, Kaiser, & Sies, 1989; Stahl & Sies, 1996). Several genes are responsible for the high content of fruit lycopene (Hp-1, hp-2, dg, and ogc) and carotenoids (Liu, Gur, & Ronen, 2003; Stevens & Rick, 1986).

Tomato fruit yield is a complex characteristic that is associated with various genetic and nongenetic factors, making it difficult to pass on to offspring. Genetic markers of yield such as QTL were found in different interspecific populations of tomatoes and were mapped on all 12 chromosomes. Genes responsible for fruit ripening in tomatoes, such as the *rin* gene, have been identified and characterized, their loci have been mapped, and their effects on tomato fruit ripening have been traced (Fox & Giovannoni, 2007; Giovannoni, Yen, & Shelton, 1999; Giovannoni, 2001).

17.7 Tomato genetic markers for selection

The use of classical morphological markers in the selection of tomatoes is associated with difficulties such as expression of dominance, epistatic interactions, and pleiotropic effects. More than 1300 morphological, physiological, and disease resistance genes have been found in the tomato population, including those for sterility, fruit ripening, and isozyme genes mapped to the 12 tomato chromosomes (Kalloo, 1991; Tanksley & Bernatzky, 1987; Tanksley, 1993).

With the advent of modern molecular DNA markers, many limitations associated with morphological and isozyme markers have been avoided by significantly advancing the effectiveness of plant genetic and breeding research. A molecular DNA marker is a small region of DNA that shows sequence polymorphism between individuals within or between species. DNA markers allow scanning of the entire genome. DNA markers can be obtained using various methods such as RFLP, RAPD, AFLP, variable number of tandem repeats, SSR, CAPS, sequence-characterized amplified regions, single-strand conformation polymorphisms, expressed sequence tags, conserved ortholog sets, and SNPs (Adams, Kelley, & Gocayne, 1991; Botstein, White, Skolnick, & Davis, 1980; Fulton, van der Hoeven, Eannetta, & Tanksley, 2002; Jeffreys, Wilson, & Thein, 1985; Konieczny & Ausubel, 1993; Orita, Iwahana, Kanazawa, Hayashi, & Sekiya, 1989; Paran & Michelmore, 1993; Tautz, 1989; He, Poysa, & Yu, 2003; Williams, Kubelik, Livak, Rafalski, & Tingey, 1990).

The use of PCR-based markers in tomato selection is increasing because they are easier to use, cheaper, faster, and less time-consuming to develop compared to the use of RFLP and AFLP (Huang, Cui, Weng, Zabel, & Lindhout, 2000; Zhang & Stommel, 2001). One of the problems in the development of markers in tomatoes is the lack of polymorphism in cultivated species or between cultivated species (Ruiz, García-Martínez, Picó, Gao, & Quiros, 2005; Williams & St. Clair, 1993). This limits the use of genetic markers in breeding programs that attempt to use intraspecific genetic variations. One way to overcome this limitation is to use high-resolution genetic markers such as SNPs that detect polymorphisms between individuals or varieties of tomatoes of one species (Labate & Baldo, 2005; Suliman-Pollatschek, Kashkush, Shats, Hillel, & Lavi, 2002; Yang, Bai, & Kabelka, 2004).

Marker-assisted selection (MAS) for any tomato trait requires precise information about gene loci and molecular markers (Francia et al., 2005; Kumari, Mir, Tyagi, Balyan, & Gupta, 2019). MAS as a method of genotyping based on SNPs is used for the selection of disease-resistant varieties (Foolad, 2007a, 2007b; Sonah et al., 2013). Thus millions of SNPs have been identified in different parts of the genome (Tomato Genome Sequencing Consortium et al., 2014). Technological advances have created several branches of omics that include genomics, transcriptomics, proteomics, metabolomics, phenomics, and ionomics, thus providing a comprehensive study of processes at different structural levels (Chaudhary, Deshmukh, Mir, & Bhat, 2019; Fukushima, Kusano, Redestig, Arita, & Saito, 2009; Hong, Yang, Zhang, & Shi, 2016; Shah et al., 2018). The development of new methods of DNA sequencing gives a significant impetus to the development of genomics and transcriptomics of species, while proteomics and metabolomics remain less developed. Tomatoes with their great economic importance and commercial value require the integration of different scientific achievements from different fields to create quality varieties with high yields in all adverse environmental conditions.

17.8 Genomic selection for abiotic stress in tomato

Genomic selection is effective in simultaneously tracking all loci contributing to the development of a trait regardless of size, making it one of the best methods for predicting genetic selection traits using molecular markers in combination with population phenotypic traits (Shah et al., 2018). Genomic selection is used to improve the yield, weight, and taste

of tomatoes. The use of phenotyping together with genomic information helps to improve the accuracy of prediction and accelerate the desired genetic gains by shortening the reproductive cycle (Yamamoto et al., 2016).

17.9 Tomato transcriptomics

In response to various adverse environmental factors, the tomato plant activates protective mechanisms for their elimination. Understanding these regulatory cascades is important for effective control of abiotic stressors. Comparing the transcripts of different tissues and at different stages of development contributes to a deeper understanding of the effective regulation of the protective responses of tomatoes to environmental stressors and to the identification of the genes involved in these mechanisms (Shinde, Behpour, McElwain, & Ng, 2015). Microchips have been used to differentially detect gene transcription in response to various abiotic stresses, including salinity, cold, drought, and oxidative stress. With the rapid development of next-generation sequencing, RNA sequencing is becoming a very economical, efficient, and high-performance transcriptome technology. Thus the method is not only limited to comparing transcript levels but is also effective in detecting new genes. The information generated by the use of microchips is extremely useful for finding regulatory genes and molecular mechanisms for drought tolerance in tomatoes (Albert et al., 2018; Iovieno et al., 2016).

17.10 Tomato proteomics

About 52 proteins have been identified in tomato leaves in response to stress from overwetting. They are involved in various energy and metabolic processes such as photosynthesis, disease resistance, stress, and defense mechanisms (Ahsan et al., 2007). Tolerance to cold damage is due to the prevention of protein denaturation and the activation of antioxidant compounds (Salazar-Salas et al., 2017). Proteins responsive to NaCl, NaHCO₃, temperature stress, and drought have been identified (Gong et al., 2014; Muneer, Ko, Wei, Chen, & Jeong, 2016; Tamburino et al., 2017). Using 2 D-gel electrophoresis and MALDI-TOF/TOF Ms, 67 proteins were detected in tomato seedlings in response to high temperature. New and more in-depth proteomic studies will help identify more candidate proteins for the development of stress-resistant and higher yielding and quality tomato varieties (Sang et al., 2017).

17.11 Tomato metabolomics

Metabolomics is a method that provides a biochemical assessment of the phenotype of an organism by identifying and quantifying low molecular weight molecules that are closely related to important toxicological and nutritional characteristics. Genomics, transcriptomics, and proteomics are not sufficient to fully and completely identify cellular mechanisms. Therefore it is necessary to study the primary and secondary metabolites. Methods such as gas chromatography–mass spectrometry, capillary electrophoresis–mass spectrometry, and nuclear magnetic resonance were used to determine metabolites in tomato plants as a protective response to stress (Kaspar et al., 2011; Lee, Perdian, Song, Yeung, & Nikolau, 2012; Schripsema, 2010). To determine the water stress on the production of flavonoids, different varieties of drought-resistant tomatoes were studied. In five varieties of cherry tomatoes, water stress leads to a reduction in shikimate and phenolic compounds (Ampofo-Asiama et al., 2014). Storage of fruits in a closed atmosphere causes low oxygen stress, which changes the metabolic profile of tomato cells through the accumulation of glycolysis intermediates, lactate, and sugar alcohols (Ampofo-Asiama et al., 2014). The integration of metabolomics, genomics, and transcriptomics provides comprehensive information on natural variations in metabolism, its genetic and biochemical control in tomatoes, and the development of tolerant tomato plants with increased yields (Zhu et al., 2018). A total of 42 positive and 76 negative mQTLs have been identified that are involved in the regulation of carbon and nitrogen metabolism in tomato leaves in response to stressors (Nunes-Nesi et al., 2019).

References

- Adams, M. D., Kelley, J. M., Gocayne, J. D., et al. (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science (New York, N.Y.)*, 252(5013), 1651–1656.
- Ahsan, N., Lee, D. G., Lee, S. H., Kang, K. Y., Bahk, J. D., Choi, M. S., . . . Lee, B. H. (2007). A comparative proteomic analysis of tomato leaves in response to waterlogging stress. *Physiologia Plantarum*, 131, 555–570.

- Albert, E., Duboscq, R., Latreille, M., Santoni, S., Beukers, M., Bouchet, J. P., Bitton, F., Gricourt, J., Poncet, C., Gautier, V., et al. (2018). Allele-specific expression and genetic determinants of transcriptomic variations in response to mild water deficit in tomato. *The Plant Journal: For Cell and Molecular Biology*, *96*, 635–650.
- Alseekh, S., Tohge, T., Wendenberg, R., Scossa, F., Omranian, N., Li, J., et al. (2015). Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. *The Plant Cell*, *27*(3), 485–512.
- Ampofo-Asiama, J., Baiye, V., Hertog, M., Waelkens, E., Geeraerd, A., & Nicolai, B. J. P. B. (2014). The metabolic response of cultured tomato cells to low oxygen stress. *Plant Biology*, *16*, 594–606.
- Anderson, L. K., Covey, P. A., Larsen, L. R., Bedinger, P., & Stack, S. M. (2010). Structural differences in chromosomes distinguish species in the tomato clade. *Cytogenetic and Genome Research*, *129*(1–3), 24–34.
- Andolfo, G., Sanseverino, W., Rombauts, S., Van der Peer, J., Bradeen, J. M., Carpato, D., ... Ercolano, M. R. (2013). Overview of tomato (*Solanum lycopersicum*) candidate pathogen recognition genes reveals important Solanum R locus dynamics. *The New Phytologist*, *197*(1), 223–237.
- Asins, J., Bretó, M. P., Cambra, M., & Carbonell, E. A. (1993). Salt tolerance in *Lycopersicon* species. I. Character definition and changes in gene expression. *Theoretical and Applied Genetics*, *86*(6), 737–743.
- Ballester, A.-R., Tikunov, Y., Molthoff, J., Grandillo, S., Viquez-Zamora, M., de Vos, R., et al. (2016). Identification of loci affecting accumulation of secondary metabolites in tomato fruit of a *Solanum lycopersicum* × *Solanum chmielewskii* introgression line population. *Frontiers in Plant Science*, *7*, 1428. Available from <https://doi.org/10.3389/fpls.2016.01428>.
- Barceloux, D. G. (2009). Potatoes, tomatoes, and solanine toxicity (*Solanum tuberosum* L., *Solanum lycopersicum* L.). *Disease-a-Month*, *55*(6), 391–402.
- Barrantes, W., Fernández-del-Carmen, A., López-Casado, G., González-Sánchez, M. A., Fernán-dez-Muñoz, R., Granell, A., & Monforte, A. J. (2014). Highly efficient genomics-assisted development of a library of introgression lines of *Solanum pimpinellifolium*. *Molecular Breeding*, *34*(40), 1817–1831.
- Barrantes, W., López-Casado, G., García-Martínez, S., Alonso, A., Rubio, F., Ruiz, J. J., et al. (2016). Exploring new alleles involved in tomato fruit quality in an introgression line library of *Solanum pimpinellifolium*. *Frontiers in Plant Science*, *7*. Available from <https://doi.org/10.3389/fpls.2016.01172>, art. 1172.
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–331.
- Bretó, M. P., Asins, M. J., & Carbonell, E. A. (1993). Genetic variability in *Lycopersicon* species and their genetic relationships. *Theoretical and Applied Genetics*, *86*(1), 113–120.
- Bulgarian Farmer. (2016). Cherry tomatoes are among the most famous and used. (online, in BG). <http://www.bgfermer.bg/Article/551128>.
- Calafiore, R., Ruggieri, V., Raiola, A., Rigano, M. M., Sacco, A., Hassan, M. I., et al. (2016). Exploiting genomics resources to identify candidate genes underlying antioxidants content in tomato fruit. *Frontiers in Plant Science*, *7*. Available from <https://doi.org/10.3389/fpls.2016.00397>, art. 397.
- Chaudhary, J., Deshmukh, R., Mir, Z. A., & Bhat, J. A. (2019). *Metabolomics: An emerging technology for soybean improvement. Biotechnology products in everyday life* (pp. 175–186). Berlin, Germany: Springer.
- Cooper, R., & Nicola, G. (2014). *Natural products chemistry: Sources, separations and structures*. CRC Press.
- Doganlar, S., Frary, A., Daunay, M. C., Lester, R. N., & Tanksley, S. D. (2002). A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the Solanaceae. *Genetics*, *161*, 1697–1711.
- Duvick, D. N. (1986). Plant breeding: Past achievements and expectations for the future. *Economic Botany*, *40*(3), 289–297.
- Duvick, D. N. (1996). Personal perspective plant breeding, an evolutionary concept. *Crop Science*, *36*(3), 539–548.
- Eklund, T. (1985). The effect of sorbic acid and esters of p-hydroxybenzoic acid on the protonmotive force in *Escherichia coli* membrane vesicles. *Journal of General Microbiology*, *131*, 73–76.
- El-Awady, M. A. M., El-Tarras, A. A. E., & Hassan, M. M. (2012). Genetic diversity and DNA fingerprint study in tomato (*Solanum lycopersicum* L.) cultivars grown in Egypt using simple sequence repeats (SSR) markers. *African Journal of Biotechnology*, *11*(96), 16233–16240.
- Eltayeb, E. A., & Roddick, J. G. (1984). Changes in the alkaloid content of developing fruits of tomato (*Lycopersicon esculentum* Mill.). II. Effects of artificial acceleration and retardation of ripening. *Journal of Experimental Botany*, *35*, 261–267.
- Ercolano, M. R., Sanseverino, W., Carli, P., Ferriello, F., & Frusciantè, L. (2012). Genetic and genomic approaches for R-gene mediated disease resistance in tomato: Retrospects and prospects. *Plant Cell Reports*, *31*, 973–985.
- FAOSTAT. (2018). (online). <http://faostat3.fao.org/browse/Q/QC/E>.
- Farrar, R. J., & Kennedy, G. G. (1991). Insect and mite resistance in tomato. In G. Kalloo (Ed.), *Genetic improvement of tomato* (pp. 122–141). Berlin, Germany: Springer, Vol. 14 of monographs on theoretical and applied genetics.
- Farrar, R. R. J., Barbour, J. D., & Kennedy, G. G. (1994). Field evaluation of insect resistance in a wild tomato and its effects on insect parasitoids. *Entomologia Experimentalis et Applicata*, *71*(3), 211–226.
- Foolad, M. R. (1999). Comparison of salt tolerance during seed germination and vegetative growth in tomato by QTL mapping. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, *42*(4), 727–734.
- Foolad, M. R. (2005). Breeding for abiotic stress tolerances in tomato. In M. Ashraf, & P. J. C. Harris (Eds.), *Abiotic stresses: Plant resistance through breeding and molecular approaches* (pp. 613–684). New York, NY: The Haworth Press.
- Foolad, M. R. (2007a). Genome mapping and molecular breeding of tomato. *International Journal of Plant Genomics*, *2007*, 52. Available from <https://doi.org/10.1155/2007/64358>, 64358.

- Foolad, M. R. (2007b). *Current status of breeding tomatoes for salt and drought tolerance* (pp. 669–700). Berlin, Germany: Springer Science and Business Media LLC.
- Foolad, M. R., & Chen, F. Q. (1998). RAPD markers associated with salt tolerance in an interspecific cross of tomato (*Lycopersicon esculentum* × *L. pennellii*). *Plant Cell Reports*, 17(4), 306–312.
- Foolad, M. R., Chen, F. Q., & Lin, G. Y. (1998). RFLP mapping of QTLs conferring cold tolerance during seed germination in an interspecific cross of tomato. *Molecular Breeding*, 4(6), 519–529.
- Foolad, M. R., & Jones, R. A. (1991). Genetic analysis of salt tolerance during germination in *Lycopersicon*. *Theoretical and Applied Genetics*, 81(3), 321–326.
- Foolad, M. R., Stoltz, T., Dervinis, C., Rodriguez, R. L., & Jones, R. A. (1997). Mapping QTLs conferring salt tolerance during germination in tomato by selective genotyping. *Molecular Breeding*, 3(4), 269–277.
- Fox, E., & Giovannoni, J. (2007). Genetic control of fruit ripening. In M. K. Razdan, & A. K. Mattoo (Eds.), *Genetic improvement of solanaceous crops* (pp. 343–378). Enfield, NH: Science Publishers.
- Framar. (2016). (online). <http://hranene.framar.bg/>.
- Francia, E., Tacconi, G., Crosatti, C., Barabaschi, D., Bulgarelli, D., Dall’Aglia, E., & Valè, G. (2005). Marker assisted selection in crop plants. *Plant Cell, Tissue and Organ Culture*, 82, 317–342.
- Frary, A., Nesbitt, T. C., Frary, A., Grandillo, S., van der Knaap, E., Cong, B., et al. (2000). fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science (New York, N.Y.)*, 289(5476), 85–88.
- Fukushima, A., Kusano, M., Redestig, H., Arita, M., & Saito, K. (2009). Integrated omics approaches in plant systems biology. *Current Opinion in Chemical Biology*, 13, 532–538.
- Fulton, M., van der Hoeven, R., Eannetta, N. T., & Tanksley, S. D. (2002). Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *The Plant Cell*, 14(7), 1457–1467.
- Gerster. (1997). The potential role of lycopene for human health. *Journal of the American College of Nutrition*, 16(2), 109–126.
- Gerszberg, A., Hnatuszko-Konka, K., Kowalczyk, T., & Kononowicz, A. K. (2015). Tomato (*Solanum lycopersicum* L.) in the service of biotechnology. *Plant Cell, Tissue and Organ Culture*, 120(3), 881–902.
- Ghosh, D., & Konishi, T. (2007). Anthocyanins and anthocyanin-rich extracts: Role in diabetes and eye function. *Asia Pacific Journal of Clinical Nutrition*, 16(2), 200–208.
- Giovannoni, J. J. (2001). Molecular biology of fruit maturation and ripening. *Annual Review of Plant Physiology and Plant Molecular Biology*, 52, 725–749.
- Giovannoni, J. J., Yen, H., Shelton, B., et al. (1999). Genetic mapping of ripening and ethylene-related loci in tomato. *Theoretical and Applied Genetics*, 98(6–7), 1005–1013.
- Gong, B., Zhang, C., Li, X., Wen, D., Wang, S., Shi, Q., & Wang, X. (2014). Identification of NaCl and NaHCO₃ stress responsive proteins in tomato roots using iTRAQ-based analysis. *Biochemical and Biophysical Research Communications*, 446, 417–422.
- Grandillo, S., Ku, H.-M., & Tanksley, S. D. (1996). Characterization of fs8.1, a major QTL influencing fruit shape in tomato. *Molecular Breeding*, 2(3), 251–260.
- Hanson, P. M., Bernacchi, D., Green, S., Tanksley, S. D., Muniyappa, V., Padmaja, A. S., . . . Chen, J. (2000). Mapping a wild tomato introgression associated with Tomato yellow leaf curl virus resistance in a cultivated tomato line. *Journal of the American Society for Horticultural Science*, 125, 15–20.
- Hanson, P. M., Sitathani, K., Sadashiva, A. T., Yang, R.-Y., Graham, E., & Ledesma, D. (2007). Performance of *Solanum habrochaites* LA1777 introgression line hybrids for marketable tomato fruit yield in Asia. *Euphytica*, 158(1–2), 167–178.
- Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., DiMeo, J., Efcavitch, J. W., et al. (2008). *Science (New York, N.Y.)*, 320, 106–109.
- HealthAliciousNess. (2016). Top 10 foods highest in lycopene. From: U.S. Agricultural Research Service Nutrition Data Releases. (online). <https://www.healthaliciousness.com/articles/high-lycopene-foods.php>.
- He, C., Poysa, V., & Yu, K. (2003). Development and characterization of simple sequence repeat (SSR) markers and their use in determining relationships among *Lycopersicon esculentum* cultivars. *Theor Appl Gen*, 106(2), 363–373.
- Heiting, G. (2014). *Lutein and zeaxanthin: Eye and vision benefits. All about vision*. Access Media Group (online). Available from <http://www.allaboutvision.com/nutrition/lutein.htm>.
- Hong, J., Yang, L., Zhang, D., & Shi, J. (2016). Plant metabolomics: An indispensable system biology tool for plant science. *International Journal of Molecular Sciences*, 17, 767.
- Huang, C., Cui, Y.-Y., Weng, C. R., Zabel, P., & Lindhout, P. (2000). Development of diagnostic PCR markers closely linked to the tomato powdery mildew resistance gene Ol-1 on chromosome 6 of tomato. *Theoretical and Applied Genetics*, 101(5–6), 918–924.
- Iovieno, P., Punzo, P., Guida, G., Mistretta, C., Van Oosten, M. J., Nurcato, R., Bostan, H., Colantuono, C., Costa, A., Bagnaresi, P., et al. (2016). Transcriptomic changes drive physiological responses to progressive drought stress and rehydration in tomato. *Frontiers in Plant Science*, 7, 371.
- Jeffreys, J., Wilson, V., & Thein, S. L. (1985). Hypervariable ‘minisatellite’ regions in human DNA. *Nature*, 314(6006), 67–73.
- Jenkins, J. A. (1948). The origin of the cultivated tomato. *Economic Botany*, 2, 379–392.
- Jones, A., & Qualset, C. O. (1984). Breeding crops for environmental stress tolerance. In G. B. Collins, & J. F. Petolino (Eds.), *Application of genetic engineering to crop improvement* (pp. 305–340). Dordrecht, The Netherlands: Nijhoff/Junk.
- Juvik, A., Berlinger, M. J., Ben-David, T., & Rudich, J. (1982). Resistance among accessions of the genera *Lycopersicon* and *Solanum* to four of the main insect pests of tomato in Israel. *Phytoparasitica*, 10, 145–156.

- Kaloo, G. (1991). *Genetic improvement of tomato*. Berlin, Germany: Springer.
- Karp, A. (2002). The new genetic era: Will it help us in managing genetic diversity? In J. M. M. Engels, V. Ramanatha Rao, A. H. D. Brown, & M. T. Jackson (Eds.), *Managing plant genetic diversity* (pp. 43–56). Wallingford: CABI Publishing. Available from <http://doi.org/10.1079/97808519952.29.0043>.
- Kaspar, S., Peukert, M., Mock, H. P., Svatos, A., Matros, A., & Mock, H. (2011). MALDI-imaging mass spectrometry—An emerging technique in plant biology. *Proteomics*, *11*, 1840–1850.
- Katsumata, A., Kimura, M., Saigo, H., Aburaya, K., Nakano, M., Ikeda, T., . . . Nagai, R. (2011). Changes in esculoside A content in different regions of the tomato fruit during maturation and heat processing. *Journal of Agricultural Food Chemistry*, *59*(8), 4104–4110.
- Khush, S., & Rick, C. M. (1968). Cytogenetic analysis of the tomato genome by means of induced deficiencies. *Chromosoma*, *23*, 452–484.
- Kong, K. W., Khoo, H. E., Prasad, K. N., Ismail, A., Tan, C. P., & Rajab, N. F. (2010). Revealing the power of the natural red pigment lycopene. *Molecules (Basel, Switzerland)*, *15*, 959–987.
- Konieczny, A., & Ausubel, F. M. (1993). A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. *Plant Journal*, *4*(2), 403–410.
- Koo, D.-H., Jo, S.-H., Bang, J.-W., Park, H.-M., Lee, S., & Choi, D. (2008). Integration of cytogenetic and genetic linkage maps unveils the physical architecture of tomato chromosome 2. *Genetics*, *179*(3), 1211–1220.
- Koushan, K., Rusovici, R., Li, W., Ferguson, L. R., & Chalam, K. V. (2013). The role of lutein in eye-related disease. *Nutrients*, *5*, 1823–1839.
- Kozukue, N., Han, J. S., Lee, K. R., & Friedman, M. (2004). Dehydrotomatine and alpha-tomatine content in tomato fruits and vegetative plant tissues. *Journal of Agricultural and Food Chemistry*, *52*(7), 2079–2083.
- Krishna, J., Bhaumik, A., & Kumar, P. (2013). Phytochemical analysis and antimicrobial studies of various extracts of tomato (*Solanum lycopersicum* L.). *Scholars Academic Journal of Biosciences*, *1*(2), 34–48.
- Ku, H.-M., Doganlar, S., Chen, K.-Y., & Tanksley, S. D. (1999). The genetic basis of pear-shaped tomato fruit. *Theoretical and Applied Genetics*, *99*(5), 844–850.
- Ku, H.-M., Grandillo, S., & Tanksley, S. D. (2000). fs8.1, a major QTL, sets the pattern of tomato carpel shape well before anthesis. *Theoretical and Applied Genetics*, *101*(5–6), 873–878.
- Kumari, S., Mir, R. R., Tyagi, S., Balyan, H. S., & Gupta, P. K. (2019). Validation of QTL for grain weight using MAS-derived pairs of NILs in bread wheat (*Triticum aestivum* L.). *Journal of Plant Biochemistry and Biotechnology*, *28*, 336–344.
- Labate, J. A., & Baldo, A. M. (2005). Tomato SNP discovery by EST mining and resequencing. *Molecular Breeding*, *16*(4), 343–349.
- Labate, J. A., Robertson, L. D., & Baldo, A. M. (2009). Multilocus sequence data reveal extensive departures from equilibrium in domesticated tomato (*Solanum lycopersicum* L.). *Heredity*, *103*(3), 257–267.
- Lee, Y. J., Perdian, D. C., Song, Z., Yeung, E. S., & Nikolau, B. J. (2012). Use of mass spectrometry for imaging metabolites in plants. *The Plant Journal: For Cell and Molecular Biology*, *70*, 81–95.
- Liu, Y.-S., Gur, A., Ronen, G., et al. (2003). There is more to tomato fruit colour than candidate carotenoid genes. *Plant Biotechnology Journal*, *1*(3), 195–207.
- Manabe, H., Murakami, Y., El-Aasr, M., Ikeda, T., Fujiwara, Y., Ono, M., & Nohara, T. (2011). Content variations of the tomato saponin esculoside A in various processed tomatoes. *Journal of Natural Medicines*, *65*(1), 176–179.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bembien, L. A., Berka, J., Braveman, M. S., Chen, Y. J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*, 376–380.
- Martin, G. B., Williams, J. G., & Tanksley, S. D. (1991). Rapid identification of markers linked to a Pseudomonas resistance gene in tomato by using random primers and near-isogenic lines. *Proceedings of the National Academy of Sciences of the USA*, *88*, 2336–2340.
- Martinez-Valverde, I., Periago, M., Provan, G., & Chesson, A. (2002). Phenolic compounds, lycopene and antioxidant activity in commercial varieties of tomato (*Lycopersicon esculentum*). *Journal of the Science of Food and Agriculture*, *82*, 323–330.
- Mascio, D., Kaiser, S., & Sies, H. (1989). Lycopene as the most efficient biological carotenoid singlet oxygen quencher. *Archives of Biochemistry and Biophysics*, *274*(2), 532–538.
- McGee, H. (July 29, 2009). Accused, yes, but probably not a killer. *The New York Times*. Retrieved 03.11.16.
- McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*, 1527–1541.
- Miller, J. C., & Tanksley, S. D. (1990). RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. *Theoretical and Applied Genetics*, *80*(4), 437–448.
- Moco, S., Capanoglu, E., Tikunov, Y., Bino, R. J., Boyacioglu, D., Hall, R. D., . . . De Vos, C. H. (2007). Tissue specialization at the metabolite level is perceived during the development of tomato fruit. *Journal of Experimental Botany*, *58*, 4131–4146.
- Muigai, S. G., Bassett, M. J., Schuster, D. J., & Scott, J. W. (2003). Greenhouse and field screening of wild *Lycopersicon germplasm* for resistance to the whitefly *Bemisia argentifolii*. *Phytoparasitica*, *31*(1), 27–38.
- Muneer, S., Ko, C. H., Wei, H., Chen, Y., & Jeong, B. R. (2016). Physiological and proteomic investigations to study the response of tomato graft unions under temperature stress. *PLoS One*, *11*, e0157439.
- Mutschler, A. R., Doerge, W., Liu, S.-C., Kuai, J. P., Liedl, B. E., & Shapiro, J. A. (1996). QTL analysis of pest resistance in the wild tomato *Lycopersicon pennellii*: QTLs controlling acylsugar level and composition. *Theoretical and Applied Genetics*, *92*(6), 709–718.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, *320*, 1344–1349.

- Nice, K. (2013). Antimicrobial screening of secondary metabolites from *Solanaceae*. Ph.D. thesis, Royal Holloway, University of London.
- Nunes-Nesi, A., Alseekh, S., Silva, F. M. D. O., Omranian, N., Lichtenstein, G., Mirnezhad, M., González, R. R. R., Garcia, J. S. Y., Conte, M., Leiss, K. A., et al. (2019). Identification and characterization of metabolite quantitative trait loci in tomato leaves and comparison with those reported for fruits and seeds. *Metabolomics: Official Journal of the Metabolomic Society*, 15, 46.
- Omodamiro, O. D., & Amechi, U. (2013). The phytochemical content, antioxidant, antimicrobial and anti-inflammatory activities of *Lycopersicon esculentum* (Tomato). *Asian Journal of Plant Science Research*, 3(5), 70–81.
- Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K., & Sekiya, T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, 86(8), 2766–2770.
- Paran, I., & Michelmore, R. W. (1993). Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. *Theoretical and Applied Genetics*, 85(8), 985–993.
- Peralta, I. E., & Spooner, D. M. (2007). History, origin and early cultivation of tomato (*Solanaceae*). In M. K. Razdan, & A. K. Mattoo (Eds.), *Genetic improvement of Solanaceous crops. Volume 2: Tomato*. USA: Science Publishers.
- Peters, S. A., van Haarst, J. C., Jesse, T. P., Woltinge, D., Jansen, K., Hesselink, T., . . . Klein-Lankhorst, R. M. (2006). TOPAAS, a tomato and potato assembly assistance system for selection and finishing of bacterial artificial chromosomes. *Plant Physiology*, 140, 805–817.
- Peterson, D. F., Stack, S. M., Price, H. J., & Johnston, J. S. (1996). DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 39(1), 77–82.
- Peterson, D. G., Pearson, W. R., & Stack, S. M. (1998). Characterization of the tomato (*Lycopersicon esculentum*) genome using in vitro and in situ DNA reassociation. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 41(3), 346–356.
- Riaz, M., Zia-Ul-Haq, M., & Saad, B. (2016). Chapter 7. The role of anthocyanins in health as antioxidant, in bone health and as heart protecting agent. *Anthocyanins and human health: Biomolecular and therapeutic aspects* (pp. 87–91). Springer.
- Rick, C. M. (1973). Potential genetic resources in tomato species: Clues from observations in native habitats. In A. M. Srb (Ed.), *Genes, enzymes, and populations* (pp. 255–269). New York, NY: Plenum Press.
- Rick, C. M. (1976a). Natural variability in wild species of *Lycopersicon* and its bearing on tomato breeding. *Genet Agraria*, 30, 249–259.
- Rick, C. M. (1976b). Tomato, *Lycopersicon esculentum* (*Solanaceae*). In N. W. Simmonds (Ed.), *Evolution of crop plants* (pp. 268–273). London, UK: Longman.
- Rick, C. M. (1978). The tomato. *Science American*, 23, 76–87.
- Rick, C. M. (1979). Potential improvement of tomatoes by controlled introgression of genes from wild species. In *Proceedings of the conference on broadening the genetic base of crops* (pp. 167–173). Pudoc, Wageningen, The Netherlands, July 1979.
- Rick, C. M. (1988). Tomato-like nightshades: Affinities, autoecology and breeders' opportunities. *Economic Botany*, 42, 145–154.
- Rick, C. M. (1991). Tomato paste: A concentrated review of genetic highlights from the beginnings to the advent of molecular genetics. *Genetics*, 128(1), 1–5.
- Rick, C. M., & Butler, L. (1956). Cytogenetics of the tomato. *Advances in Genetics*, 8, 267–382.
- Rick, C. M., DeVerna, J. W., & Chetelat, R. T. (1990). Experimental introgression to the cultivated tomato from related wild nightshades. In A. B. Bennett, & S. D. O'Neill (Eds.), *Horticultural biotechnology symposium* (pp. 19–30). Davis, CA: John Wiley & Sons.
- Rick, C. M., DeVerna, J. W., Chetelat, R. T., & Stevens, M. A. (1987). Potential contributions of wide crosses to improvement of processing tomatoes. *Acta Horticulturae*, 200, 45–55.
- Rick, C. M., & Fobes, J. F. (1975). Allozyme variation in the cultivated tomato and closely related species. *Bulletin of the Torrey Botanical Club*, 102(6), 376–384.
- Rick, C. M., & Holle, M. (1990). Andean *Lycopersicon esculentum* var. *cerasiforme*: Genetic variation and its evolutionary significance. *Economic Botany*, 43(Suppl. 3), 69–78.
- Rossi, M., Goggin, F. L., Milligan, S. B., Kaloshian, I., Ullman, D. E., & Williamson, V. M. (1998). The nematode resistance gene Mi of tomato confers resistance against the potato aphid. *Proceedings of the National Academy of Sciences of the USA*, 95(17), 9750–9754.
- Ruiz, J. J., García-Martínez, S., Picó, B., Gao, M., & Quiros, C. F. (2005). Genetic variability and relationship of closely related Spanish traditional cultivars of tomato as detected by SRAP and SSR markers. *Journal of the American Society for Horticultural Science*, 130(1), 88–94.
- Salazar-Salas, N. Y., Valenzuela-Ponce, L., Vega-García, M. O., Pineda-Hidalgo, K. V., Vega-Alvarez, M., Chavez-Ontiveros, J., . . . Lopez-Valenzuela, J. A. (2017). Protein changes associated with chilling tolerance in tomato fruit with hot water pre-treatment. *Postharvest Biology and Technology*, 134, 22–30.
- Sang, Q., Shan, X., An, Y., Shu, S., Sun, J., & Guo, S. (2017). Proteomic analysis reveals the positive effect of exogenous spermidine in tomato seedlings' response to high-temperature stress. *Frontiers in Plant Science*, 8, 555.
- Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., et al. (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 15, 227–239.
- Schalk, J. M., & Stoner, A. K. (1976). A bioassay differentiates resistance to the Colorado potato beetle on tomatoes. *Journal of the American Society for Horticultural Science*, 101, 74–76.
- Schripsema, J. (2010). Application of NMR in plant metabolomics: Techniques, problems and prospects. *Phytochemical Analysis*, 21, 14–21.
- Scott, J. W. (1993). Breeding tomatoes for resistance to high temperatures, biotic and abiotic diseases. *Hort Bras*, 11, 167–170.
- Scott, J. W., Bryan, H. H., & Ramos, L. J. (1997). High temperature fruit setting ability of large-fruited, jointless pedicel tomato hybrids with various combinations of heat-tolerance. *Proceedings of the Florida State Horticultural Society*, 110, 281–284.

- Scott, J. W., Everett, P. H., Bryan, H. H., et al. (1985). Suncoast—A large-fruited home garden tomato. *Florida Agricultural Experiment Stations Circular*, S-322.
- Scott, J. W., Francis, D. M., Miller, S. A., Somodi, G. C., & Jones, J. B. (2003). Tomato bacterial spot resistance derived from PI 114490; inheritance of resistance to race T2 and relationship across three pathogen races. *Journal of the American Society for Horticultural Science*, 128(5), 698–703.
- Scott, J. W., & Jones, J. B. (1986). Sources of resistance to bacterial spot in tomato. *HortScience: A Publication of the American Society for Horticultural Science*, 21, 304–306.
- Scott, J. W., Jones, J. B., & Somodi, G. C. (1995). Screening tomato accessions for resistance to *Xanthomonas campestris* pv. *vesicatoria*, race T3. *HortScience: A Publication of the American Society for Horticultural Science*, 30, 579–581.
- Scott, J. W., Miller, S. A., Stall, R. E., et al. (1997). Resistance to race T2 of the bacterial spot pathogen in tomato. *HortScience: A Publication of the American Society for Horticultural Science*, 32(4), 724–727.
- Scott, J. W., Olson, S. M., Howe, T. K., Stoffella, P. J., Bartz, J. A., & Bryan, H. H. (1995). ‘Equinox’ heat-tolerant hybrid tomato. *HortScience: A Publication of the American Society for Horticultural Science*, 30, 647–648.
- Scott, J. W., Volin, R. B., Bryan, H. H., & Olson, S. M. (1986). Use of hybrids to develop heat tolerant tomato cultivars. *Proceedings of the Florida State Horticultural Society*, 99, 311–315.
- Shah, T., Xu, J., Zou, X., Cheng, Y., Nasir, M., & Zhang, X. (2018). Omics approaches for engineering wheat production under abiotic stresses. *International Journal of Molecular Sciences*, 19, 2390.
- Shinde, S., Behpouri, A., McElwain, J. C., & Ng, C. K. Y. (2015). Genome-wide transcriptomic analysis of the effects of sub-ambient atmospheric oxygen and elevated atmospheric carbon dioxide levels on gametophytes of the moss, *Physcomitrella patens*. *Journal of Experimental Botany*, 66, 4001–4012.
- Shirasawa, K., Isobe, S., Hirakawa, H., Asamizu, E., Fukuoka, H., Just, D., et al. (2010). SNP discovery and linkage map construction in cultivated tomato. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 17(6), 381–391.
- Sies, H., & Stahl, W. (1995). Vitamins E and C, B-carotene and other carotenoids as antioxidants. *American Journal of Clinical Nutrition*, 65, 1315–1321.
- Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., Normandeau, E., Laroche, J., LaRose, S., Jean, M., et al. (2013). An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One*, 8, e54603.
- Stahl, W., & Sies, H. (1996). Lycopene: A biologically important carotenoid for humans? *Archives of Biochemistry and Biophysics*, 336(1), 1–9.
- Stankovic, I. (2004). Zeaxanthin. Chemical and technical assessment. FAO 63rd JECFA.
- Stevens, M. A., & Rick, C. M. (1986). Genetics and breeding. In J. G. Atherton, & J. Rudich (Eds.), *The tomato crop: A scientific basis for improvement* (pp. 35–109). New York, NY: Chapman and Hall.
- Stevens, M. R., Lamb, E. M., & Rhoads, D. D. (1995). Mapping the Sw-5 locus for tomato spotted wilt virus resistance in tomatoes using RAPD and RFLP analyses. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 90, 451–456.
- Stommel, J. R., & Zhang, Y. P. (1998). Molecular markers linked to quantitative trait loci for anthracnose resistance in tomato. *HortScience: A Publication of the American Society for Horticultural Science*, 33, 514.
- Suliman-Pollatschek, S., Kashkush, K., Shats, H., Hillel, J., & Lavi, U. (2002). Generation and mapping of AFLP, SSRs and SNPs in *Lycopersicon esculentum*. *Cellular and Molecular Biology Letters*, 7(2A), 583–597.
- Tamburino, R., Vitale, M., Ruggiero, A., Sassi, M., Sannino, L., Arena, S., Costa, A., Batelli, G., Zambrano, N., Scaloni, A., et al. (2017). Chloroplast proteome response to drought stress and recovery in tomato (*Solanum lycopersicum* L.). *BMC Plant Biology*, 17, 40.
- Tanksley, D., & Bernatzky, R. (1987). Molecular markers for the nuclear genome of tomato. In D. J. Nevins, & R. A. Jones (Eds.), *Plant biology, vol. 4, tomato biotechnology* (pp. 37–44). New York, NY: Alan R. Liss.
- Tanksley, S., & Fulton, T. (2007). Dissecting quantitative trait variation—examples from the tomato. *Euphytica*, 154, 365–370.
- Tanksley, S. D. (1993). Linkage map of the tomato (*Lycopersicon esculentum*) (2N = 24). In S. J. O’Brian (Ed.), *Genetic maps: Locus maps of complex genomes* (pp. 6.3–6.15). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, 17(16), 6463–6471.
- The Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400), 635–641.
- Tigchelaar, E. C. (1986). Tomato breeding. In M. J. Bassett (Ed.), *Breeding for vegetable crops* (pp. 135–171). Westport, CO: AVI.
- Tomato Genome Sequencing Consortium., Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., et al. (2014). Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *The Plant Journal: For Cell and Molecular Biology*, 80, 136–148.
- USDA Food Composition Databases. (2016). Tomato. United States Department of Agriculture Agricultural Research Service. (online) <https://ndb.nal.usda.gov/ndb/search/list?qlookup=11529&format=Full>.
- van der Hoeven, R. S., Ronning, C., Giovannoni, J. J., Martin, G., & Tanksley, S. D. (2002). Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *The Plant Cell*, 14(7), 1441–1456.
- Varietal Seeds and Plant Protection Ltd. (2016). (online) <http://sortovisemena-bg.com/>.
- Villard, J., Skroch, P. W., Lai, T., Hanson, P., Kuo, C. G., & Nienhuis, J. (1998). Genetic variation among tomato accessions from primary and secondary centers of diversity. *Crop Science*, 38, 1339–1347.

- Wang, Y., van der Hoeven, R., Nielsen, R., et al. (2006). Characteristics of the tomato nuclear genome as determined by sequencing unmethylated DNA and euchromatic and pericentromeric BACs. In *Plant and animal genome XIV conference* (p. 147). USDA, San Diego, CA, January 2006, Abstract W176.
- Wang, Y., van der Hoeven, R. S., Nielsen, R., Mueller, L. A., & Tanksley, S. D. (2005). Characteristics of the tomato nuclear genome as determined by sequencing undermethylated EcoRI digested fragments. *Theoretical and Applied Genetics*, *112*(1), 72–84.
- Warnock, S. J. (1988). A review of taxonomy and phylogeny of the genus *Lycopersicon*. *HortScience: A Publication of the American Society for Horticultural Science*, *23*(4), 669–673.
- Warren, G. F. (1998). Spectacular increases in crop yields in the United States in the twentieth century. *Weed Technology*, *12*(4), 752–760.
- Webb, D. (2014). Anthocyanins. *Today's Dietitian*, *16*(3), 20.
- Weston, A., Johnson, D. A., Burton, H. T., & Snyder, J. C. (1989). Trichome secretion composition, trichome densities, and spider mite resistance of ten accessions of *Lycopersicon hirsutum*. *Journal of the American Society for Horticultural Science*, *114*(3), 492–498.
- Williams, C. E., & St. Clair, D. A. (1993). Phenetic relationships and levels of variability detected by restriction fragment length polymorphism and random amplified polymorphic DNA analysis of cultivated and wild accessions of *Lycopersicon esculentum*. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, *36*(3), 619–630.
- Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, *18*(22), 6531–6535.
- Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., Shirasawa, K., Hirakawa, H., Nunome, T., Yamaguchi, H., et al. (2016). A simulation-based breeding design that uses whole-genome prediction in tomato. *Scientific Reports*, *6*, 19454.
- Yamamoto, T., Yoshimura, M., Yamaguchi, F., Kouchi, T., Tsuji, R., Saito, M., . . . Kikuchi, M. (2004). Anti-allergic activity of naringenin chalcone from a tomato skin extract. *Biosci Biotechnol Biochem*, *68*(8), 1706–1711.
- Yang, W., Bai, X., Kabelka, E., et al. (2004). Discovery of singly nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. *Molecular Breeding*, *14*(1), 21–34.
- Yang, W., Sacks, E. J., Lewis, M. L. I., Miller, S. A., & Francis, D. M. (2005). Resistance in *Lycopersicon esculentum* intraspecific crosses to race T1 strains of *Xanthomonas campestris* pv. *vesicatoria* causing bacterial spot of tomato. *Phytopathology*, *95*(5), 519–527.
- Young, A., & MacArthur, J. W. (1947). Horticultural characters of tomatoes, Texas Agricultural Experiment Station Bulletin No 698, 1947.
- Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. *Nature Reviews. Genetics*, *2*(12), 983–989.
- Zhang, Y., & Stommel, J. R. (2001). Development of SCAR and CAPS markers linked to the Betas gene in tomato. *Crop Science*, *41*(5), 1602–1608.
- Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., et al. (2018). Rewiring of the fruit metabolome in tomato breeding. *Cell*, *172*, 249–261.

Characterization of drought tolerance in maize: omics approaches

Ramandeep Kaur¹, Manjot Kaur¹, Parampreet Kaur² and Priti Sharma¹

¹School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India, ²School of Organic Farming, Punjab Agricultural University, Ludhiana, India

18.1 Introduction

Maize is the third most significant crop for food, feed, and necessary raw material for different industries after wheat and rice. It belongs to the *Poaceae* family and is a tall annual plant with an extensive fibrous root framework. Maize is cultivated in every country across the world with a total production of 1099.61 million tons (Statista, 2020). The United States has been the major producer of maize and it is the driver of the US economy, trailed by China, representing about 40% and 25%, respectively. In India, area under maize cultivation is 9,027,130 ha with 27,715,100 tonnes production (Food & Agriculture Organisation of the United Nations FAOSTAT, 2019). About 15 Million farmers are engaged in maize cultivation and it provides employment to more than 650 million people in India. The distinctive characteristics of maize make the crop a suitable crop candidate for enhancing farmer's income and livelihoods in India. States, for example, Karnataka, Rajasthan, Andhra Pradesh, and Madhya Pradesh, contribute toward half of the total maize acreage in the nation. Though, it is extremely pertinent to observe that the national efficiency of maize is considerably lesser than the global standards. India stands almost half the global yield standards and therefore there lies immense scope for improvement as the strategically important crop in the country. A shift in the global demand of cereals, particularly maize requirements of developing countries to surpass that of rice and wheat by 2020 was already projected as reported in Pingali and Heisey (2001). More than 85% of the maize land is possessed under rainfed conditions during the monsoon pattern and rising temperature prompts different abiotic constraints which thus contribute an apparent decrease in yield profitability (Sheikh, 2017). Rise in temperature and change in climatic conditions leads to many different abiotic constraints. Amongst the all abiotic stresses, drought stress is considered as a most devastating environmental stress amongst the natural anxieties worldwide as it decreases yield profitability (Sheikh, 2017) and has rendered enormous region of global agricultural land inefficient or unproductive (Huang et al., 2015; Langridge & Reynolds, 2015; Obidiegwu, Bryan, Jones, & Prashar, 2015; Zhan, Schneider, & Lynch, 2015).

Drought is a multidimensional stress imposing a series of cellular processes, such as morphological, physiological, and biochemical to get adapted to dehydration. Leaf rolling, stomatal closure, membrane stability, osmotic alteration, antioxidant accumulation, reactive oxygen species (ROS) are some of the manifestation of adaptation for drought condition. Besides it reduces leaf size, stem expansion and root multiplication, disturbs plant water relations and diminishes water-use productivity, which ultimately hampers crop productivity to extreme levels. The general outline about drought tolerance is given in Fig. 18.1.

Maize inflorescence (male and female flowers) is most susceptible to drought stress throughout the flowering time (Grant, Jackson, Kiniry, & Arkin, 1989; Pantuwan, Fukai, Cooper, Rajatasereekul, & O Toole, 2002). Timing of drought is imperative as if the stress occurs before flowering, the silk growth with respect to male flowering gets delayed and results in a prolonged anthesis-silking interval (ASI) (Bolanos & Edmeades, 1996). The extended ASI causes the silks to all dried up when the pollen reaches it during fertilization (Bassetti & Westgate, 1994) or may be after the period when ovaries have used their saved starch (Saini & Westgate, 2000; Zinselmeier, Habben, Westgate, & Boyer, 2000). This results in hindered ear and silk development causing kernel and ear abortion (Edmeades, Bolanos, &

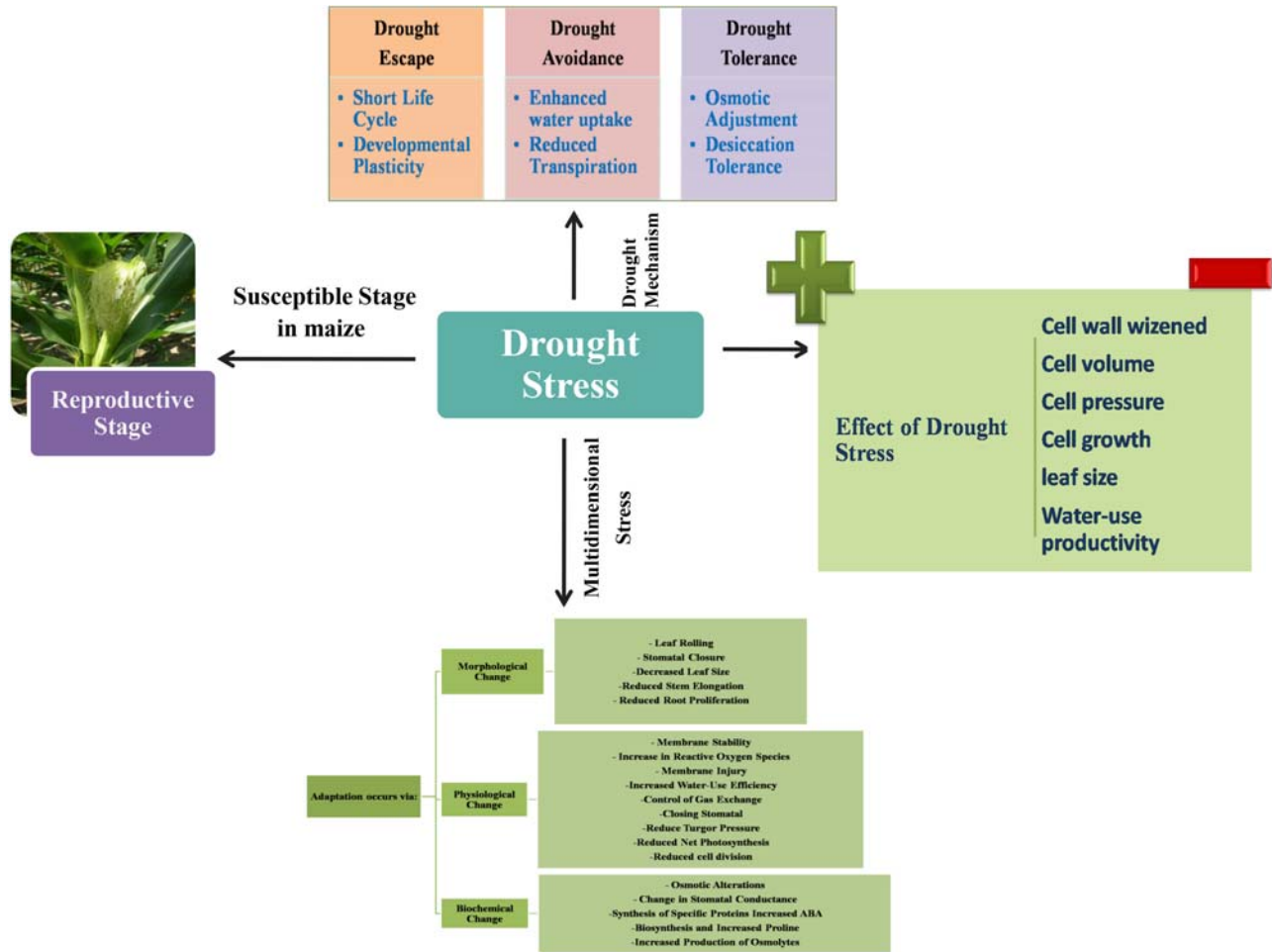


FIGURE 18.1 Schematic representation of drought tolerance in plants.

Lafitte, 1992) and eventually loss in productivity. Maize productivity is reported to get reduced by 15%–30% (Lobell et al., 2014) following increased water scarcity as a function of climatic drifts.

As per survey by Living-water Ltd, United Kingdom (2018), India is listed at second position among the drought prone countries after Morocco. Therefore there is an urgent need to make concrete efforts to accomplish the increasing demand for food for highly populated geographical areas with water scarcity. Current scenario makes it imperative to understand the response and adaptation of maize to water/drought stress. Generally, plant drought resistance involves drought escape via a short life cycle or developmental plasticity (Manavalan, Guttikonda, Phan Tran, & Nguyen, 2009), drought avoidance via enhanced water uptake and reduced water loss (Luo, 2010; Tardieu, 2013), or drought tolerance via osmotic adjustment, antioxidant capacity, and desiccation tolerance (Luo, 2010; Yue et al., 2006).

18.2 Drought timing

Depending upon the time point drought occurs in a life cycle of plant, it could be either terminal drought and intermittent drought. In terminal drought, the soil water availability decreases slowly and this condition leads to rigorous drought stress at the later grain filling and development stages. Intermittent drought occurs as a consequence of limited periods of insufficient rain or irrigation occurring at one or more interval throughout the growing season and is not necessarily lethal. Stages of drought, that is, in premature stage, mid-season, or terminal stage transpires the effect on productivity and quality losses. In drought prone areas, strategies used to maintain crop yield which means, breeding varieties with improved yield under drought stress as well as under irrigated conditions. Early stage drought stress reduces plant growth and inhibits plant development (Heiniger, 2001; Shaw, 1983). Drought that occurs at V8 (Vegetative stage) (Collar of 8th leaf visible occurs approximately 45 days after emergence) to V17 (8 weeks after

emergence) leaf stages has considerable impact on plant architecture, growth, cob size and number of kernels (Farre & Faci, 2006; Westgate & Boyer, 1985). Drought stress that occurs between 2 weeks before or after the emergence of silk could be the reason of significant decreases in overall revenue in terms of number of kernel's set and their weight (Schussler & Westgate, 1991; Westgate & Bassetti, 1990), resulting in an average yield loss of 20%–50% (ACIAR; Nielsen, 2007).

18.3 Plant response to drought

Growth maintenance of crop under drought is more substantial than its survival (Dolferus, 2014). So for this maintenance, different types of drought-adaptive strategies/mechanisms are evolved in plants, which allow them to acclimatize to specific environmental conditions for their growth and development (Fang & Xiong, 2015). Various morphological and physiological responses are involved in plants under drought stress for adaptation (Wang & Huang, 2004) as shown in Fig. 18.2. Several physiological factors are affected due to drought stress injury such as drought stress damages both photosynthetic apparatus and reduce chlorophyll content (Jiang & Huang, 2001). An in-depth insight into the mechanisms of drought response and adaptation adopted by the plants is essential for the accomplishment of the goals to develop drought tolerant crop varieties. Numerous changes occurring at physiological and developmental levels in plant as a response to stress conditions are governed by the expression of stress inducible genes (Philippe et al., 2010). These



FIGURE 18.2 Effect of drought stress on morphophysiological traits.

genes further interact with numerous partners, and constitutes a complex interactome governing a drought tolerance response (Lin et al., 2007). Thus plant response to drought stress depends upon the genetic makeup of plant, plant species, plant age, the severity of drought and developmental stage (Ali, Basra, Munir, Mahmood, & Yousaf, 2011; Gall et al., 2015).

Agronomists defined the term Drought resistance in terms of “relative yield of genotypes” or “the capacity of a plant to produce an economic product with minimum loss in a water-scarcity environment in comparative to the water-constraint free management” (Fang & Xiong, 2015; Fukai & Cooper, 1995). Drought resistance is a complex trait and its expression depends on the action and interaction of several morphological, physiological, and biochemical features. (May & Milthorpe, 1962) reported three types of drought resistance, that is, *drought escape*, *drought avoidance* and *drought tolerance*.

Plants regulate their growth period to avoid moisture stress is labeled as *drought escape*. Under favorable conditions, plants undergo longer vegetative phase because shorter vegetative period reduces the time available photosynthetic productivity and accumulation of seed nutrients resulting in an overall decline in plant biomass. This mechanism, involves several key factors such as early flowering and early maturity (phenology), developmental plasticity as well as remobilization of assimilates (preanthesis) to developing grain (Turner, 1979). The plants thus usually undergo a short vegetative period. This is a classical adaptive mechanism in which plants undergo speedy development in order to complete the full life-cycle of growth before onset of drought stress. It is observed in some cereal crops such as maize. In maize, speedy flowering time and shorter vegetative period in response to terminal drought can be very important and beneficial. During the sensitive period of flowering, the above mentioned strategies in maize can help to reduce exposure to dehydration and plants evolve postanthesis grain filling. The Eastern part of India and Bangladesh are drought-prone upland areas, in these areas, drought escape is an imperative phenomenon which permits maize to produce kernels even under restricted water accessibility conditions (Bernier, Atlin, Serraj, Kumar, & Spaner, 2008). *Drought avoidance* is the sustaining of important physiological processes such as stomatal regulation, when exposed to mild drought. In this mechanism, relatively high tissue water potential is maintained regardless of deficiency of soil moisture. Water stress avoidance strategy is adopted by all those maize varieties in which plants maintain their water status even under stress conditions through their well-developed roots organization and structure. In such varieties, yield losses thus minimized caused by drought as plants store water in cells with less water loss (Singh, van Oosterom, Jordan, & Hammer, 2012). Maize varieties that avoid drought generally have deep roots, higher root:shoot, increased root surface, branching rate, and root length with a higher penetrability, high cuticular resistance, stomatal closure in the early phase and suppleness in leaf rolling (Wang et al., 2006). *Drought tolerance* is the ability of plants to produce a higher yield at limited tissue water content. In cereals, drought tolerance mechanism usually works in the reproductive phase. Tolerant cultivars show better germination, seedling growth, and photosynthesis. It involves amalgamation of several mechanisms aiming at avoiding or tolerating water scarcities and relies on the capacity of plants to undergo severe dehydration via osmotic alteration and osmoprotectants. *Drought recovery* is defined as the plants ability to recommence growth and gain yield after plant gets exposure to rigorous drought stress which causes absolute loss of turgor pressure and dehydration of leaves (Fang & Xiong, 2015; Luo, 2010). Recovery after stress is a very multifaceted process as it involves the reorganization of many metabolic pathways to restore drought-induced damage and to continue plant growth again. It requires far more than simply a return to the state before stress onset (Vankova, Dobra, & Storchova, 2012).

18.4 Progress with conventional breeding strategies for drought tolerance in maize

Breeding programs are used for screening, characterizing, and identifying germplasm for drought tolerance for transferring of the required trait into the elite varieties so as to maximize the yield gains under water stress conditions. Deep root system in maize genotypes has been observed as an important feature to tolerate the drought effects on its growth and yield. Haseeb, Nawaz, Rao, Ali, and Malik (2020) observed two genotypes B-316 and Raka-poshi with improved performance for all traits involved in drought tolerance, particularly shoot and root length, thus suggesting that selection of maize genotypes on the basis of root length, shoot length and dry shoot weight may be fruitful for development of drought stress tolerance maize hybrids and synthetic varieties. Guo et al. (2020) studied the phenotype associated with the root morphological characters and estimated the drought tolerance index using different maize association panel at seedling stage and concluded that the seminal root length could be beneficial for improving drought tolerance.

Djemel et al. (2018) evaluated 51 different open-pollinated maize populations from diverse temperate regions under water deficit conditions at germination, seedling establishment and early growth stage and identified potential sources of drought tolerance which could be used to study and get a deeper insight into drought tolerance under different development stages. They concluded BS17 maize population to exhibit higher germination rate, early vigor, and rapid

seedling growth, no effect on water use efficiency under drought conditions. Another population, Enano Levantino/Hembrilla exhibited no effect on the stomatal conductance, rate of transpiration, and photosynthesis. Similar studies thus hold the potential to offer new possibilities for development of drought tolerant hybrids using breeding programs by combining diverse mechanisms conferring tolerance through crossings between potential donors (Djemel et al., 2018).

18.4.1 Seedling and physiological traits for drought tolerance

Trait-based approaches are often used to predict the ecological consequences of climate change. Relationships of seedling and root traits are more commonly measured traits. Higher genotypic coefficient of variation was observed for different traits, that is, dry root weight and fresh shoot weight (Mehdi & Ahsan, 2000), fresh shoot weight, fresh root weight, dry root weight and dry shoot weight (Mehdi & Ahsan, 2000), dry shoot weight, dry root weight, emergence percentage, fresh shoot weight and fresh seedling weight (Khan, Habib, Sadaqat, & Tabir, 2004) suggesting that these traits can be used as selection criteria while selecting families for high green maize fodder yield. Ahsan et al. (2011) found that fresh shoot length and fresh root weight were positively correlated with fresh shoot weight. The partial dominance effect was observed for thermostability of cell membrane, net photosynthetic rate (Chohan, Muhammad, & Muhammad, 2012) and mean germination time increases with decrease in osmotic potential (Khodarahmpour, 2012) under drought stress conditions. Positive correlations were found between shoot and root traits with medium to high heritability of shoot and root seedling traits (Badr, El-Shazly, Tarawneh, & Börner, 2020). Several studies have been conducted through selection of physiological traits to identify a candidate drought resistant maize genotypes with higher yield generally includes cell membrane thermostability, stomatal conductance, survival rate of maize seedlings (Aslam, Iftikhar, Saleem, & Ali, 2006), length and biomass of root, root density, shoot biomass, and leaf temperature, etc. (Ali et al, 2011).

18.4.2 Yield traits for drought tolerance

Ultimately, yield productivity under stress conditions is a prime factor in selection of a drought tolerant genotype. Significant positive genotypic correlation was found for plant height with grain yield (GY)/plant, number of cobs/plant, grains/cobs and 100-seed weight (Ali et al., 2011; Banziger & Diallo, 2004; Khatun, Begham, Motin, Yasmine, & Islam, 1999; Umakanth, Satyanarayana, & Kumar, 2000; Waseem et al., 2014) and these can be used as selection criteria for improving GY. Also, maize GY was found to be positively and significantly associated with circumference and diameter of cob, and also with number of grain rows per cob (Vaezi, Mishani, Samadi, & Ghannadhs, 2000). From correlation analysis, it was concluded that grains per cob, cobs per plant, cob length, and 100-seed weight significantly affects the GY (Torun & Koycu, 1999). The general and specific combining ability effects were observed to be significantly high for all yield related parameters, that is, cobs per plant, height of plant, grain rows per cob, weight of 100 seeds, leaf area GY per plant, length and weight of cob (Akbar & Saleem, 2008; Gautam, 2003; Muraya, Ndirangu, & Omolo, 2006; Zhou, Cheng, Yaohal, & Young, 2004). But in some studies, general combining ability effects were significantly positive for grain quality traits and other agronomic traits except GY (Bhatnagar, Betran, & Rooney, 2004). Under drought stress, the significant genetic variation, heritability and superior performance of several quantitative traits could help to increase GY (Qayyum et al., 2003). Hader (2006) stated that leaf area may be a convenient indirect factor to improve maize yield, through a positive and direct effect and significant correlation with size of stomata and frequency (Ahsan, Hadar, Saleem, & Aslam, 2008). Significant positive association between stomatal conductance with GY and flag leaf area could be used as benchmarks for the selection of higher yielding maize genotypes (Yousafzai, Al-Kaff, & Moore, 2009) Further, morphophysiological traits, that is, flag leaf area, cobs per plant, green fodder yield, cob length and weight, grain rows per cob, plant height, GY per plant and grain weight (100 seeds) can be used as selection criterion for the development of higher yield drought tolerant maize variety and fodder yield hybrid (Saif-ul-malook, Ali, Shakeel, Sajjad, & Bashir, 2014).

Conventional plant breeding approaches have been used to address the drought tolerance potential in maize (Richards, 1985), however plant response based selection criterion are affected by low heritability, genetic interaction, environment and genotype interactions, and polygenic effects, thus making the selection process time-consuming and laborious as immense phenotypic screening is mandatory. Traditional approaches needs to integrated with technological advancements so as to provide a required momentum for providing drought tolerant maize cultivars and varieties.

18.5 Omics for characterizing drought stress responses in maize

Modern technological advancements have led to the development of high performance tools which can be utilized to delve into and examine the plant genomes for crop betterment. In this aspect, “Omics” approaches have risen with most promising perspectives of developing varieties with improved quality. The omics methodologies work to decipher the whole genome to look up into plant molecular responses and these high-throughput and integrated approaches have been used in several crops successfully to investigate the temporal and spatial system fluctuation that occurs under various stresses.

Maize response to drought stress involves rearrangements occurring during gene expression at molecular level, beginning from transcription regulation, leading to mRNA processing, translation and modification. Under stress conditions, the transcriptome expression is affected due to plants stress specific regulation of transcription. Some of the functions of these transcriptionally regulated genes are to mediate transcription, translation, signaling, metabolism, and common stress responses. Generally, stages such as vegetative and reproductive have been observed to be more inclined toward stress. In recent years, ample amount of survey has been undertaken to deduce the mechanism involved in numerous stress tolerance in crop plants. Modernistic progressions in microarray and intense deep sequencing technologies have led to augmentation of genomic and transcriptomic data under several abiotic stresses. To date, transcript levels are commonly used as the only measure for gene upregulation/downregulation in high throughput approaches also known as expression analysis which is stage specific. A considerable amount of studies has shown a depressed correlation among the transcripts and protein content starting the imperative post transcriptional processes as defined predictive value of transcripts for protein expression.

18.5.1 Genomics

In recent years, the potential and advancement in DNA sequencing technologies (2nd/3rd generation sequencers) have been attributed to abundance of sequence information giving rise to whole genome sequences providing an in depth peek through in the physical structure of genome. The maize genetics and genomics database (Maize GDB) is the model database for maize. It currently hosts 12 fully sequenced and assembled maize genomes, including B73, B104, CML247, W22, Mo17, PH207, EP1, and F7. The current B73 assembly version, Zm-B73-REFERENCE-NAM-5.0 released in January 2020, was sequenced and assembled along with a set of 25 inbred lines known as the NAM founder lines by the NAM Consortium. These advances have given birth to the exploitation of plant genomics studies for breeding climate resilient varieties utilizing gene/ quantitative trait loci (QTL) identification, SNP marker development, etc (Table 18.1). Usage of QTL (Capelle, Remoue, & Moreau, 2010) and oligomicroarray (Luo, 2010) scrutinizes, various QTLs/genes linked with kernel desiccation were found to be involved in ABA synthesis. Various studies have reported that many drought responsive proteins or genes, such as *ZmTPA*, *ZmRFPI*, and *ZmCPK4* were induced by ABA dependent or ABA independent manner. Studies by Huang, Møller, and Song (2012); Jiang, Zhang, and Wang (2013); and Xia, Liu, Wu, and Ding (2012) identified major effect QTLs on chromosome (Chr) number 1, 2, 8 and 10 for maize drought tolerance in a set of 230 recombinant inbred lines developed by CIMMYT. A major QTL for ASI (anthesis-silking interval) and ear number per plant under drought stress was detected on Chr 1 (bin 1.03) and Chr 9 (bins 9.03–9.05) (Hao et al., 2008) from a cross between X178 and B73 which corresponded to several QTLs identified in different experiments carried out worldwide. Several such identified “consensus QTLs” have served as good candidates in marker assisted breeding program to enhance maize production under drought stress. MARS (Marker Assisted Recurrent Selection) has been used to improve frequency of favorable alleles in maize through bi-parental population that combined drought tolerance with resistance against armyworm infestation (*Striga hermonthica*) (Abdulmalik et al., 2017). Bankole et al. (2017) used MARS approach in maize and suggested the effectiveness of this method to improve drought tolerance and GY in a biparental population developed from drought tolerant lines. Zhou, Dong, and Shi (2017) identified a major QTL associated with grain weight using simple sequence RBackspace repeats (SSR) markers and further, *qGW1.05*-NILs were developed by Marker Assisted Selection. Cerrudo et al. (2018) also used QTL-MAS approach to improve genetic gain for tolerance to drought in maize (Table 18.1).

Functional genomics: The introductory knowledge on molecular phenotypes exposes genotypic variation that dominates morphophysiological traits. Functional genomic studies have been proven to be the most pertinent knowledge for crop improvement. Functional genomics give the applicability to study gene functions and interactions between genes and their regulatory network. These systems can be exploited to generate improved crop varieties using either sequence or hybridization based technologies.

TABLE 18.1 List of genes/ quantitative trait loci associated with drought stress.

Genes/quantitative trait loci (QTLs)	Chromosome no.	Annotation	Reference
5 QTLs	1,2,3,5	Anthesis-silking interval (ASI)	Li et al. (2003)
3 QTLs	2, 6		Zhang et al. (2004)
43 QTLs	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Grain yield, leaf width, plant height, ear height, leaf number	Nikolic A et al. (2011)
11 QTLs	1, 2, 3, 5, 6, 7	ASI, ear setting	Xin-Hai and Xian-DE (2003)
22 QTLs	1, 3, 6, 7, 9	Sugar concentration, relative leaf water, root density, root dry weight, total biomass, grain yield, osmotic potential	Rahman et al. (2011)
12 QTLs	2, 3, 4, 5, 7, 10	Kernel per ear, grain yield	Nikolic et al. (2013)
64 QTLs	1, 2, 3, 4, 5, 6, 10	Relative leaf water content	Nikolic et al. (2012)
9 QTLs	3, 6, 7, 8	ASI	Liu et al. (2010)
34 QTLs			Chen et al. (2012)
9 QTLs	1, 2, 3, 4, 5, 6, 7	Cob weight	Upadayaula et al. (2006)
3 QTLs	1, 7	SPAD	Trachsel et al. (2010)
1 QTL	1	Chlorophyll content	Messmer et al. (2011)
1 QTL	3	Chlorophyll content	Almeida et al. (2014)
1 QTL	9	Ear weight	Zhao et al. (2018)
3 QTLs	1, 7	Kernel number per ear	Ribaut et al. (1997)
6 QTLs	1, 3, 9	Kernel per ear	Xiao et al. (2005)
9 QTLs	1, 4, 7, 8, 9	Grain yield	Cerrudo et al. (2018)
18 QTL	3, 5, 7, 10	Grain yield, ear setting, ASI	Hu et al. (2021)
827 probe sets	–	Differential expression levels of cell-wall related and transporter genes	Zheng et al. (2010)
619 Genes and 126 transcripts	–	Altered regulation under drought conditions. Beta-amylase, chitinase, carotenoid hydroxylase, heat shock proteins and were upregulated by drought.	Song et al. (2017)
29 Differentially expressed proteins	–	Involved in metabolism, stress response, photosynthesis, and protein modification	Kim et al. (2015)
61 Differentially expressed proteins	–		Zhao et al. (2016)
1 Gene	–	<i>bZIP4</i>	Ma et al. (2018)
1 Gene	–	<i>bZIP17</i>	Jia et al. (2009)
1 Gene	–	<i>DREB2.7</i>	Liu et al. (2013)
2 Gene	–	<i>DBF1, DBF2</i>	Kizis and Pagès (2002)

(Continued)

TABLE 18.1 (Continued)

Genes/quantitative trait loci (QTLs)	Chromosome no.	Annotation	Reference
1 Gene	–	<i>NAC111</i>	Mao et al. (2015)
1 Gene	–	<i>NF-YA1</i>	Luan et al. (2015)
1 Gene	–	<i>NF-YA3</i>	Su et al. (2018)
1 Gene	–	<i>NF-YB2</i>	Nelson et al. (2007)
1 Gene	–	<i>NF-YB16</i>	Wang et al. (2018)
1 Gene	9	ZmTIP1—encodes S-acyltransferase	Zhang et al. (2020)
4552 Differentially expressed genes	–	Phenylpropanoid biosynthesis, taurine metabolism and cell wall biosynthesis.	Zhang et al. (2020)

Sequence based approaches: Expressed sequence tags (ESTs) are one of the earliest methods to study gene and genome annotation. Over millions of EST data have been deposited at National Center for Biotechnology Information database. Around 2 million ESTs are drawn from 10 inbred lines are deposited in GenBank. EST sequencing can be utilized extensively even if the entire transcriptome is not fully represented. EST sequencing has the potential for gene discovery by comparing different genotypes under both controlled and stressed conditions. A computer-based methodology was developed by [Batley, Barker, O’Sullivan, Edwards, and Edwards \(2003\)](#) to aid in the identification of candidate single nucleotide polymorphisms (SNPs) as well as small insertions/deletions from expressed sequence tag data. The applicability of this method was verified by applying these SNPs and insertions/deletions to 102,551 maize (*Zea mays*) expressed sequence tag sequences screening out a total of 14,832 candidate polymorphisms. The predicted SNPs and insertion/deletions represent true genetic variation in maize. [Hao et al. \(2011\)](#) identified 1536 SNP markers using Illumina GoldenGate assay and genotyped maize inbred lines. Furthermore, they determined the functional genetic variations underlying drought tolerance by association analysis. A total of 1006 polymorphic SNPs were detected. Pairwise linkage disequilibrium and association mapping with phenotypic traits was done under water stress and irrigated conditions and about 29 SNPs were found to be affiliated with two phenotypic traits which were correlated with drought tolerant genes. Another approaches, that is, Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing expression data for maize is publicly available (<https://mpss.udel.edu>) which can be interpreted and analyzed for gene expression studies. [Poroyko et al. \(2005\)](#) made use of SAGE to characterise the relative amount of transcripts in the root tips of irrigated maize seedlings (*Z. mays* cv. FR697). A total of 161,320 tags were detected representing a minimum of 14,850 genes, based on at least two tags determined per mRNA. Moreover, they confirmed the expression of few selected transcripts correlating with tag frequency using quantitative reverse transcription-PCR.

Hybridization-based approaches, on the other hand, exploit hybridization of the perfectly matched or mismatched target DNA with the oligonucleotide or cDNA probes which are adhered to the surface to check the expression. In the array based methods, prior knowledge of the transcript is required for designing probes. In a study, [Luo \(2010\)](#) examined gene expression in developing kernels under drought stress and identified drought responsive genes. Gene expression profiles were done in the developing kernels of Tex6 maize line under both drought stress and well watered regimes using the 70-mer maize oligoarrays. About 9573, positive array spots were identified and 7988 were common under drought stress and well-watered samples. Further expression patterns of some genes in several stress response-associated pathways were examined, and it was found that specific genes were responsive to drought stress.

Since the plant genomes shared massive similarities among themselves, comparative genomics can be primarily utilized for species with unexplored genomes to access information among closely as well as distantly related plant species. Grasses are the main focus of the study via comparative genomics due to their agronomic importance. The scope of genome conservation first became distinct by genome mapping based comparative studies, which advocated a colinear order of genes and markers encompassing on genomes of different species. In spite of having comprehensive analysis which reveal notable rearrangements at molecular level such as inversions, deletions, and translocations, extensive linearity across grass genomes has been calculated for gene discovery and isolation. “Genome zipper” is one of the concepts that have emerged from comparative genomics that basically helps in determining the virtual gene order in partially sequenced genomes. The approach relies on syntenic genes. However, recently evolved genes or certain

rearrangements at the molecular level cannot be examined by this method. In contrast, nonsyntenic genes provide valuable information on evolution and speciation of the genome. To forecast the gene order and its organization in these species it is indisputable that species-specific genomic features can only be accessed through a fully annotated reference genome sequences genetically mapped markers, despite the presence of comparative genomics and genome zippers. In maize, *ZmASR3* gene, which activates the antioxidant system and regulates the ABA-dependent pathway under water stressed conditions acts as a positive regulator of drought tolerance in crop breeding programs. Its overexpression in drought stressed *Arabidopsis* showed lower malondialdehyde levels and higher relative water content and proline content than the wild type, demonstrating that *ZmASR3* can improve drought tolerance (Liang et al., 2019). Further, CDPKs (Calcium dependent protein kinases) have been shown to be involved in abiotic stress tolerance in various crop species. Over-expression of *OsCPK7* and *AtCPK6* in rice and *Arabidopsis*, respectively, led to enhanced drought tolerance (Saijo, Hata, Kyojuka, Shimamoto, & Izui, 2000; Xu et al., 2010). Mittal et al. (2017) studied the syntenic relationship among a set of 32 CDPK genes under drought stress in maize with *Arabidopsis*, rice and sorghum and led to conclusion that Maize-rice species genes were less divergent as compared to maize-sorghum and maize-*Arabidopsis* owing to the divergence during evolutionary time scale. Among transcription factors (TF), overexpression of WRKY TF – *ZmWRKY40* from maize, improved drought tolerance in transgenic *Arabidopsis* by regulating other stress responsive genes including *STZ*, *DREB2B*, and *RD29A*, and the ROS content in transgenic lines by enhancing the activities of peroxide dismutase and catalase under drought stress (Wang et al., 2018). In another study, breeding values of 240 maize subtropical lines phenotyped for drought across different environments using 29,619 cured SNPs. A total of 77 SNPs associated with 10 drought-responsive TF (stomatal closure, root development, hormonal signaling and photosynthesis) with higher marker effects was selected across all datasets to validate the genes and QTLs associated with drought tolerance (Shikha et al., 2017).

The recent enormous research done in the areas of structural, functional, and comparative genomics showed that the information generated from one plant species can be implemented for the advancement of other related species or genera or taxa.

18.5.2 Transcriptomics

Transcriptomic approaches have set the ball rolling to understand the plant responses to abiotic stresses. In past decade, application of transcriptome analysis such as next generation sequencing and RNA-seq has been used by the plant genomic resources. The ability of transcriptomic technologies to provide deep coverage and representation of abundant transcript has a huge potential. Transcriptome analysis has proven to be an advantageous tool to filter candidate genes, anticipate gene function, and detect cis-regulatory motifs. Transcriptomic changes between drought-tolerant and control maize lines were investigated using a drought-tolerant maize mutant. Delayed wilting and higher drought tolerance was observed under both controlled and field conditions in the mutant C7-2t in comparison to its wild type C7-2, entailing its high water-holding ability. A total of 4552 differentially expressed genes were identified based on transcriptomic profiling. The differentially expressed genes (DEGs) involved in phenylpropanoid biosynthesis were also enriched in cell components associated with cell wall biosynthesis and membrane systems (Zhang, Liu, Wu, & Wang, 2020).

RNA-seq has also been demonstrated to identify large-scale identification of drought-responsive genes and to expedite the extraction of key drought tolerance genes in plants, for example, maize (Lu et al., 2017). For instance, RNA-seq in maize RILs under drought stress have depicted that the upregulation of cell wall biosynthesis/aquaporin-related genes are associated with drought adaptability (Min et al., 2016). Genes conferring cell wall remodeling, biosynthesis of certain amino acids and carbohydrates have been revealed to be linked with drought-response mechanism (Zenda et al., 2019). The bioinformatics tools along with RNA-seq have been applied to study the correlation between flowering time and drought stress in maize. In a total of 619 genes identified, the expression of 126 transcripts was altered by drought stress, which included zinc finger and NAC domains. The study also identified 20 genes encoding for TF like HY5, PRR37 and CONSTANS regulating the flowering time mechanism (Song et al., 2017). Transcriptomic studies of drought-tolerant YE8112 and drought-sensitive MO17 lines identified several TF to play a primary role, that is, two WRKY genes were down-regulated, two MYB-related genes were up-regulated, and two GARP-G2-like genes were down-regulated, whilst one NAC gene was up-regulated in response to drought stress. Another study identified the role of DnaJ in conferring drought tolerance in maize, a protein belonging to heat shock protein Hsp40 family, and completing the correct folding of protein, maintain the stability of peptide chain, and prevent cell damage caused by environmental stress (Wang & Huang, 2004). Many studies (Hu et al., 2015; Thirunavukkarasu et al., 2014) have shown that DnaJ protein plays an important role in the life activities of plants to cope with environmental stress. However, there was a little effect on the progression of developmental stages during drought suppressed plant growth. The parallel

RNA-seq profiling of ears, leaves, tassels in response to drought stress conditions, at several developmental stages, exposed tissue specific differences (Danilevskaya et al., 2019). The study inferred significant down-regulation of genes which controlled DNA replication, cell cycle, and cell division in stressed ears and inflorescence meristem.

18.5.3 Proteomics and metabolomics

Drought affects multiple life processes that are convoluted in plant growth and development, such as osmotic potential adjustment, antioxidant capabilities, photosynthetic rate reduction, and abscisic acid accumulation (Cramer, Urano, Delrot, Pezzotti, & Shinozaki, 2011). Many differentially expressed proteins control these processes at various developmental stages in various resistant species. The presence of recurrent discordance between protein levels and the plenty of cognate gene transcripts proposed the demand for complementary analysis of the proteome, leading to further validation of candidate genes and pathways. There have been elaborative studies using two-dimensional gel electrophoresis (2-DE)/Mass spectrometry (Ms)-based proteomics, in seeds, leaves, and roots in response to drought in maize. Huang et al. (2012) identified several differentially expressed proteins like 17.4 kDa Class I HSP3, EMB564, and other stress responsive proteins through 2-D and Ms/MS approach, and concluded their role in conferring drought tolerance during embryogenesis and seed germination. In other study, difference was observed among the levels of drought protective proteins such as CAT, APX, and SOD in drought-tolerant and drought-sensitive maize genotypes. The drought-tolerant maize had higher accumulation of these proteins (Benesova, Hola, & Fischer, 2012). A study conducted by Bahrun, Jensen, Asch, and Mogensen (2002) on maize ABA signaling pathway, inferred that it is one of the fundamental signaling pathways that mediate maize to adapt in drought stress. Proteomic analysis of drought stressed maize led to the identification of proteins involved in metabolism, photosynthesis and stress responses (Kim et al., 2019). These key proteins were comprised of 11 metabolism-related proteins, 7 defense/stress-related proteins, 2 photosynthesis-related proteins, 1 protein involved in protein synthesis, and 3 unknown function proteins. The major class of enzymes responded to water deficit plants were related to carbohydrate metabolism. The abundance of malate dehydrogenase, two isoforms of NAD-dependent epimerase/dehydratase, alpha-galactosidase, and isocitrate dehydrogenase under water deficit conditions has been observed. Correspondingly, the revelation of proteomic analysis of Arabidopsis under drought stress showed that branched-chain amino acid aminotransferase 3 protein and zinc finger transcription factor oxidative stress 2 proteins had a vital role to play under drought stress responses in plants that over-expressed ethylene response factor AtER (Scarpeci, Frea, Zanor, & Valle, 2017). The two ABA-deficient maize mutants namely vp5 and wild-type Vp5 were analyzed under drought stress on the basis of proteomic differences using 2-DE and Ms/MS. From this analysis, it was inferred that in maize roots, proteins associated with drought stress were majorly involved in energy and metabolism, redox homeostasis, and regulatory processes (Hu et al., 2011). The proteins identified in the leaves of maize, performed a vital role in processes such as ATP synthesis, protein synthesis, chlorophyll synthesis, CO fixation, gluconeogenesis, antioxidant defense, and signal transduction. In response to light stress and drought stress, most of the proteins that differentially assembled in leaves were localized to the chloroplast, functioned in an ABA-dependent manner (Hu et al., 2012) thus suggesting the importance of ABA in regulating the synthesis of drought-induced proteins. Such studies provide an integrated picture to correlate the transcript level changes with the observed proteome structure.

Stable-isotope labeling has been done in several large scale studies to characterize maize proteome and phosphoproteome dynamics under drought stress. Bonhomme, Valot, Tardieu, and Zivy (2012) identified unique 3664 phosphorylation sites on 2496 proteins, that could potentially affect epigenetic regulation, transcriptional control, cell cycle-dependent mechanisms, phytohormone-mediated responses, histone modifications, DNA methylation patterns, and ABA-, ethylene-, auxin- and/or jasmonate-related responses through these phosphopeptidyl proteins. Benesova et al. (2012) studied and analyzed changes induced by drought in the maize leaf proteome using LC based isobaric tags for relative and absolute quantitation (iTRAQ) for proteome characterization and gel-based 2-DE analysis. They characterized 326 proteins by iTRAQ and 11 proteins by 2-DE combined with Ms/MS analysis. Out of these proteins, only four proteins were identified by both the methods. This result suggested the compatibility of the two technologies though partially overlapping. Omics studies have determined predominately protective proteins such as HSPs (Benesova et al., 2012; Hu et al., 2010; Li et al., 2009; Luo, 2010), late embryogenesis-abundant proteins (LEAs), stress response-related proteins (such as NBSLRR resistance-like protein) (Hu et al., 2012), 14-3-3-like proteins (Huang et al., 2012; Li et al., 2009), phytohormone related proteins, and signaling proteins (such as auxin repressed protein and serine/threonine protein kinase) to be involved in the maize drought response.

There is still a scope of advancement in the methodology of maize stress omics studies. For instance, most of the maize stress proteomic studies rely on gel-based technology and for protein detection methods Coomassie Brilliant

Blue Staining is used. However, the low resolution of the 2-D electrophoresis has proved to be a drawback particularly for membrane proteins. Therefore methods with increased detection based sensitivity, such as iTRAQ (gel-free), can be applied using diverse genetic backgrounds which differ greatly with an abundance of drought tolerance variations in maize stress proteomics studies.

Protein profiling provides evidence on changes in protein abundance or adjustment in response to stress that could be correlated with the metabolomics analysis to investigate the region related with a major change in levels of any metabolite. Utilizing, metabolomics approaches it has been showed that, to regulate physiological stress responses under drought stress, endogenous gene expression of ABA level rises significantly. Metabolite profiling has displayed the accumulation of ABA during drought also regulates the accumulation of various amino acids and sugars such as glucose and fructose. In particular, correlation of the drought-inducible expression of key biosynthetic genes, that is, BCAT2, LKR/SDH, P5CS1, and ADC2 with the drought-inducible accumulations of branch-chain amino acids (BCAAs), like proline, saccharopine, and agmatine respectively, can be observed, that is regulated by endogenous ABA. On the flip-side, the accumulation of galactinol and raffinose is not regulated by ABA under drought stress (Soni et al., 2015).

18.5.4 Advances in phenomics

With an available whole genome sequence of certain species, whole genome tiling arrays is one the rewarding transcript profiling to examine abiotic stress responses (Rensink & Buell, 2005). Genome wide expression profiles help in detection of candidate genes for desired traits, for example, stress tolerance. The inactivation or overexpression of these profiles will further aid in their characterization and utilization. The whole development of high-throughput phenotyping, or “phenomics,” has expanded into a highly active research field. In the area of phenomics, high throughput techniques such as robotics, spectroscopy and imaging have been introduced recently. These highly advanced technological advancements as well as high performing computing systems can efficiently analyze the obtained data (Rahaman, Chen, Gillani, Klukas, & Chen, 2015). For instance, a phenotyping system such as LemnaTec Scanalyser, it offers proximal remote sensing technology which captures the image of individual plants as well as data related to plant growth, architecture, health and responses providing help in observation and analyzing the genotype X environment interactions in precise manner (Petrozza et al., 2014). Furthermore, utilization of nondestructive imaging technologies can help in efficiently measuring the dissection of series of component trait during drought stress (Berger, Parent, & Tester, 2010). To illustrate, (Honsdorf, March, Berger, Tester, & Pillen, 2014) used a set of 47 wild barley introgression lines for the high-throughput phenotyping to evaluate drought tolerance. In this study, a significant correlation of the biomass estimated with the image processing with the actual biomass was found. High throughput phenotypic approaches have proved to be beneficial as they help in precise detection of QTLs, over the previous traditional labor-intensive measures of height, biomass, flowering time, harvest index, and GY. The root structures are vital components of drought stress tolerance. The conventional methods are highly destructive to study phenotyping of the root traits, as they involve complete removal of plants from the soil. Therefore the *era* of phenomics has catered to provide nondestructive methods for analyzing root traits through imaging under drought stress. For instance, a high throughput method (BRACE) reported by Sharma and Carena (2016) can efficiently perform phenotyping of root traits in a nondestructive manner under maize drought stress conditions. Its high speed and efficiency usually takes less than 2 min per plot. Therefore it has proven to be useful and reliable method for large-scale, high-throughput phenotyping screening. Similarly, one of the reliable, fast, and much more efficient methods such as thermography has been used for tropical maize population under water stress for high throughput phenotyping (Romano et al., 2011).

Due to the complex nature of drought stress, there has been involvement of number of genes for its study revealing the presence of multiple pathways conferring drought stress tolerance. The recent progress in genomics and bioinformatics are offering better opportunities to assess and enhance diversity in germplasm collections, introgress valuable traits from new sources and identify genes that control key traits.

18.5.5 Bioinformatics tools and databases

Rapid advances in bioinformatics has played a significant role in the development of the agricultural sector, with the increase of sequencing projects, rapid DNA and RNA sequencing tools, and high-throughput SNP genotyping techniques, functional characterization are now available to understand the genetics of drought tolerance and which are believed to play a major role in stress responses. Many publicly available databases, such as ARAMEMNON, focus on specific protein classes enables the identification of a series of stress-related integral membrane proteins in both monocots (maize, banana, rice, and brachypodium) and dicots (*Arabidopsis*, poplar, grape, tomato, and muskmelon). Alter

et al. (2015) has developed Drought Stress Gene Database, DroughtDB, a specific, valuable resource and information tool for researchers working on drought stress that includes manually curated compilation of molecularly characterized originally identified genes, their information about physiological and/or molecular function and provides detailed information about computed orthologous genes in nine model and crop plant species including maize and barley. The Plant Phosphorylation database (P³DB) enables the investigation of changes in stress-induced phosphorylated proteins in six plant species (*Arabidopsis*, rapeseed, soybean, barrelclover, rice and maize) (Zenda et al., 2019).

The coupled use of omics-data, specific genetic designs, and pertinent analytical methods provides the integrative information that increases our understanding between plant stress response and crop yield and quality. Using the bioinformatics tools, has helped researchers in finding many target genes involved in a vast diversity of functions in various plant species (Axtell & Bowman, 2008), genomic selection, rapid generation advancement, and other tools. Upadhyay et al. (2019) used psRNA Target and RNA hybrid tools to predict the target genes in maize by investigating differential expression of transcripts for drought tolerance. A total of seven microRNAs targeting 16 mRNAs were predicted and were validated by qRT-PCR. Under drought stress, the differential expression of microRNAs regulates the expression of their target genes, resulting in multiple responses of physiological and biochemical pathways relative to drought tolerance of maize (Table 18.1).

18.6 Conclusion

Numerous modern breeding techniques in conjunction with multiomics platforms along with high-throughput phenotyping have been done to identify putative QTLs/genomic regions for drought tolerance in maize. However, fine mapping of these regions and the identification of candidate genes involved in crucial metabolic pathways and mechanisms through which the associated genetic variants exert their effects are still rarely unknown. Therefore it is important to inculcate “omics” sciences into linkage and association mapping to bridge this knowledge gap. Leveraging knowledge from physiological, biochemical, and molecular regulatory mechanisms of drought response may uncover the complexity of drought tolerance and can provide a useful foundation for breeding drought-tolerant sesame hybrids. Undoubtedly, a multidisciplinary approach, will help synergistic understanding and fine-tune the development of new maize hybrids that can adapt water scarcity.

References

- Abdulmalik, R. O., Menkir, A., Meseka, S. K., Unachukwu, N., Ado, S. G., Olarewaju, J. D., ... Gedil, M. (2017). Genetic gains in grain yield of a maize population improved through marker assisted recurrent selection under stress and non-stress conditions in West Africa. *Frontiers in Plant Science*, 8, 841.
- ACIAR (Australian Centre for International Agricultural Research). n.d. *Impact of drought on corn physiology and yield*. Canberra ACT, Australia: Australian Centre for International Agricultural Research.
- Ahsan, M., Hadar, M. Z., Saleem, M., & Aslam, M. (2008). Contribution of various leaf morpho-physiological parameters towards grain yield in maize. *International Journal of Agriculture And Biology*, 10, 546–550.
- Ahsan, M., Hussain, M. M., Farooq, A., Khaliq, I., Farooq, J., Ali, Q., & Kashif, M. (2011). Physio-genetic behavior of maize seedlings at water deficit conditions. *Cercetari Agronomice in Moldova*, 44(2), 41–49.
- Akbar, M., & Saleem, M. (2008). Combining ability analysis in maize under normal temperature condition. *Journal of Agricultural Research*, 46, 39–47.
- Ali, Z., Basra, S. M. A., Munir, H., Mahmood, A., & Yousaf, S. (2011). Mitigation of drought stress in maize by natural and synthetic growth promoters. *Journal of Agriculture and Social Research*, 7(2), 56–62.
- Alter, S., Bader, K. C., Spannagl, M., Wang, Y., Bauer, E., Schön, C. C., & Mayer, K. F. (2015). DroughtDB: An expert-curated compilation of plant drought stress genes and their homologs in nine species. *Database (Oxford)*, 2015, bav046.
- Aslam, M., Iftikhar, A. K., Saleem, M., & Ali, Z. (2006). Assessment of water stress tolerance in different maize accessions at germination and early growth stage. *Pakistan Journal of Botany*, 38, 1570–1579.
- Axtell, M. J., & Bowman, J. L. (2008). Evolution of plant microRNAs and their targets. *Trends in Plant Science*, 13(7), 343–349.
- Badr, A., El-Shazly, H. H., Tarawneh, R. A., & Börner, A. (2020). Screening for drought tolerance in maize (*Zea mays L.*) germplasm using germination and seedling traits under simulated drought conditions. *Plants*, 9(5), 565.
- Bahrin, A., Jensen, C. R., Asch, F., & Mogensen, V. O. (2002). Drought-induced changes in xylem pH, ionic composition, and ABA concentration act as early signals in field-grown maize (*Zea mays L.*). *Journal of Experimental Botany*, 53(367), 251–263.
- Bankole, F., Menkir, A., Olaoye, G., Crossa, J., Hearne, S., Unachukwu, N., & Gedil, M. (2017). Genetic gains in yield and yield related traits under drought stress and favorable environments in a maize population improved using marker assisted recurrent selection. *Frontiers in Plant Science*, 8, 808.

- Banziger M., Diallo A.O. (2004). Progress in developing drought and N stress tolerant maize cultivars for eastern and southern Africa. In *Integrated approaches to higher maize productivity in the new millennium. Proceedings of the seventh Eastern and Southern Africa regional maize conference*. pp 189-94, CIMMYT/KARI, Nairobi, Kenya.
- Bassetti, P., & Westgate, M. E. (1994). Floral asynchrony and kernel set in maize quantified by image analyses. *Agronomy Journal*, 86, 699-03.
- Batley, J., Barker, G., O'Sullivan, H. J., Edwards, K., & Edwards, D. (2003). Mining for Single nucleotide polymorphisms (SNPs), insertions, and deletions can all lead to the adaptation of plants toward unfavorable environment condition. *Plant Physiology*, 132, 84–91.
- Benesova, M., Hola, D., Fischer, L., et al. (2012). The physiology and proteomics of drought tolerance in maize: Early stomatal closure as a cause of lower tolerance to short-term dehydration. *PLoS One*, 7, e38017.
- Berger, B., Parent, B., & Tester, M. (2010). High-throughput shoot imaging to study drought responses. *Journal of Experimental Botany*, 61(13), 3519–3528.
- Bernier, J., Atlin, G. N., Serraj, R., Kumar, A., & Spaner, D. (2008). Breeding upland rice for drought resistance. *Journal of the Science of Food and Agriculture*, 88(6), 927–939.
- Bhatnagar, S., Betran, F. J., & Rooney, L. W. (2004). Combining abilities of quality protein maize inbred. *Crop Science*, 44, 1997–2005.
- Bolanos, A., & Edmeades, G. O. (1996). The importance of anthesis-silking interval in breeding for drought tolerance in tropical maize. *Field Crops Research*, 48, 65–80.
- Bonhomme, L., Valot, B., Tardieu, F., & Zivy, M. (2012). Phosphoproteome dynamics upon changes in plant water status reveal early events associated with rapid growth adjustment in maize leaves. *Molecular & Cellular Proteomics*, 11(10), 957–972.
- Capelle, V., Remoue, C., Moreau, L., et al. (2010). QTLs and candidate genes for desiccation and abscisic acid content in maize kernels. *BMC Plant Biology*, 10, 2.
- Cerrudo, D., Cao, S., Yuan, Y., Martinez, C., Suarez, E. A., Babu, R., . . . Trachsel, S. (2018). Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Frontiers in Plant Science*, 9, 366.
- Chohan, M. S. M., Muhammad, S., & Muhammad, A. (2012). Genetic analysis of water stress tolerance and various morpho-physiological traits in *Zea mays* L. using graphical approach. *Pakistan Journal of Nutrition*, 11(5), 489–500.
- Cramer, G. R., Urano, K., Delrot, S., Pezzotti, M., & Shinozaki, K. (2011). Effects of abiotic stress on plants: A systems biology perspective. *BMC Plant Biology*, 11(1), 1–14.
- Danilevskaya, O. N., Yu, G., Meng, X., Xu, J., Stephenson, E., Estrada, S., . . . Thatcher, S. (2019). Developmental and transcriptional responses of maize to drought stress under field conditions. *Plant Direct*, 3(5), e00129.
- Djemel, A., Álvarez-Iglesias, L., Pedrol, N., López-Malvar, A., Ordás, A., & Revilla, P. (2018). Identification of drought tolerant populations at multi-stage growth phases in temperate maize germplasm. *Euphytica*, 214(8), 1–18.
- Dolferus, R. (2014). To grow or not to grow: A stressful decision for plants. *Plant Science (Shannon, Ireland)*, 229, 247–261.
- Edmeades G.O., Bolanos J., Lafitte H.R. (1992). Progress in breeding for drought tolerance in maize. In Wilkinson D (Ed.) *Proceedings of the fifty-seventh annual corn and sorghum industry research conference ASTA, Washington, USA*, (pp. 93–111).
- Fang, Y., & Xiong, L. (2015). General mechanisms of drought response and their application in drought resistance improvement in plants. *Cellular and Molecular Life Sciences: CMLS*, 72(4), 673–689.
- Farre, I., & Faci, J. M. (2006). Comparative response of maize (*Zea mays* L.) and sorghum (*Sorghum bicolor* L. Moench) to deficit irrigation in a Mediterranean environment. *Agricultural Water Management*, 83(1–2), 135–143.
- Food and Agriculture Organisation of the United Nations (FAOSTAT) (2019) [cited 2021]. <<http://www.fao.org/faostat/en/#data/QC>>.
- Fukai, S., & Cooper, M. (1995). Development of drought-resistant cultivars using physiomorphological traits in rice. *Field Crops Research*, 40(2), 67–86.
- Gall, H. L., Philippe, F., Domon, J. M., Gillet, F., Pelloux, J., & Rayon, C. (2015). Cell wall metabolism in response to abiotic stress. *Plants*, 4(1), 112–166.
- Gautam, A. S. (2003). Combining ability studies for grain yield and other agronomic characters in inbred lines of maize. *Journal of Crop Research*, 26, 482–485.
- Grant, R. F., Jackson, B. S., Kiniry, J. R., & Arkin, G. F. (1989). Water deficit timing effects on yield components in maize. *Agronomy Journal*, 81, 61–65.
- Guo, J., Li, C., Zhang, X., Li, Y., Zhang, D., Shi, Y., & Wang, T. (2020). Transcriptome and GWAS analyses reveal candidate gene for seminal root length of maize seedlings under drought stress. *Plant Science (Shannon, Ireland)*, 292, 110380.
- Hader. (2006). [master's thesis] *Association of various physiomorphological characters in maize (Zea mays L.)*. Pakistan: University of Agriculture Faisalabad.
- Hao, Z., Li, X., Xie, C., Li, M., Zhang, D., Bai, L., & Zhang, S. H. (2008). Two consensus quantitative trait loci clusters controlling anthesis–silking interval, ear setting and grain yield might be related with drought tolerance in maize. *Annals of Applied Biology*, 153, 73–83.
- Hao, Z., Li, X., Xie, C., Weng, J., Li, M., Zhang, D., . . . Zhang, S. (2011). Identification of functional genetic variations underlying drought tolerance in maize using SNP markers. *Journal of Integrative Plant Biology*, 53, 641–652.
- Haseeb, A., Nawaz, A., Rao, M. Q. A., Ali, Q., & Malik, A. (2020). Genetic variability and association among seedling traits of *Zea mays* under drought stress conditions. *Biological and Clinical Sciences Research Journal*, 2020(1), e020-e020.
- Heiniger, R. W. (2001). *The impact of early drought on corn yield*. Raleigh, NC: North Carolina State University.
- Honsdorf, N., March, T. J., Berger, B., Tester, M., & Pillen, K. (2014). High-throughput phenotyping to detect drought tolerance QTL in wild barley introgression lines. *PLoS One*, 9(5), e97047.

- Hu, X., Li, Y., Li, C., Yang, H., Wang, W., & Lu, M. (2010). Characterization of small heat shock proteins associated with maize tolerance to combined drought and heat stress. *Journal of Plant Growth Regulation*, 29(4), 455–464.
- Hu, X., Lu, M., Li, C., Liu, T., Wang, W., Wu, J., ... Zhang, J. (2011). Differential expression of proteins in maize roots in response to abscisic acid and drought. *Acta Physiologiae Plantarum / Polish Academy of Sciences, Committee of Plant Physiology Genetics and Breeding*, 33(6), 2437.
- Hu, X., Wu, L., Zhao, F., Zhang, D., Li, N., Zhu, G., ... Wang, W. (2015). Phosphoproteomic analysis of the response of maize leaves to drought, heat and their combination stress. *Frontiers in Plant Science*, 6, 298.
- Hu, X., Wu, X., Li, C., Lu, M., Liu, T., Wang, Y., & Wang, W. (2012). Abscisic acid refines the synthesis of chloroplast proteins in maize (*Zea mays*) in response to drought and light. *PLoS One*, 7(11), e49500.
- Huang, H., Möller, I. M., & Song, S. Q. (2012). Proteomics of desiccation tolerance during development and germination of maize embryos. *Journal of Proteomics*, 75(4), 1247–1262.
- Huang, Q., Zhao, Y., Liu, C., Zou, X., Cheng, Y., Fu, G., & Lu, G. (2015). Evaluation of and selection criteria for drought resistance in *C. hinesesemi* winter rapeseed varieties at different developmental stages. *Plant Breeding*, 134(5), 542–550.
- Jiang, S., Zhang, D., Wang, L., et al. (2013). A maize calcium dependent protein kinase gene, ZmCPK4, positively regulated abscisic acid signaling and enhanced drought stress tolerance in transgenic Arabidopsis. *Plant Physiology and Biochemistry*, 71, 112–120.
- Jiang, Y., & Huang, B. (2001). Physiological responses to heat stress alone or in combination with drought: A comparison between tall fescue and perennial ryegrass. *Horticultural Science*, 36(4), 682–686.
- Khan, I. A., Habib, S., Sadaqat, H. A., & Tabir, M. H. N. (2004). Comparative evaluation and analysis of seedling traits for drought tolerance in Maize. *Journal of Agricultural Science and Botany*, 2, 246–251.
- Khatun, F., Begham, S., Motin, A., Yasmine, S., & Islam, M. R. (1999). Correlation coefficient and path analysis of some maize hybrids. *Bangladesh Journal of Botany*, 28, 9–15.
- Khodarahmpour, Z. (2012). Evaluation of maize (*Zea mays* L.) hybrids, seed germination and seedling characters in water stress conditions. *African Journal of Agricultural Research*, 7, 6049–6059.
- Kim, S. G., Lee, J. S., Bae, H. H., Kim, J. T., Son, B. Y., Kim, S. L., ... Jeon, W. T. (2019). Physiological and proteomic analyses of Korean F1 maize (*Zea mays* L.) hybrids under water-deficit stress during flowering. *Applied Biological Chemistry*, 62(1), 1–9.
- Langridge, P., & Reynolds, M. P. (2015). Genomic tools to assist breeding for drought tolerance. *Current Opinion in Biotechnology*, 32, 130–135.
- Li, H. Y., Huang, S. H., Shi, Y. S., Song, Y. C., Zhong, Z. B., Wang, G. Y., ... Yu, L. (2009). Isolation and analysis of drought-induced genes in maize roots. *Agricultural Sciences in China*, 8(2), 129–136.
- Liang, Y., Jiang, Y., Du, M., Li, B., Chen, L., Chen, M., et al. (2019). ZmASR3 from the maize ASR gene family positively regulates drought tolerance in transgenic Arabidopsis. *International Journal of Molecular Sciences*, 20(9), 2278.
- Lin, M. K., Belanger, H., Lee, Y. J., Varkonyi-Gasic, E., Taoka, K. I., Miura, E., & Lough, T. J. (2007). FLOWERING LOCUS T protein may act as the long-distance florigenic signal in the cucurbits. *The Plant Cell*, 19(5), 1488–06.
- Lobell, D. B., Roberts, M. J., Schlenker, W., Braun, N., Little, B. B., Rejesus, R. M., & Hammer, G. L. (2014). Greater sensitivity to drought accompanies maize yield increase in the US Midwest. *Science (New York, N.Y.)*, 344(6183), 516–519.
- Lu, X., Zhou, X., Cao, Y., Zhou, M., McNeil, D., Liang, S., & Yang, C. (2017). RNA-seq Analysis of Cold and Drought Responsive Transcriptomes of *Zea mays* ssp. *mexicana* L. *Frontiers in Plant Science*, 8, 136.
- Luo, L. J. (2010). Breeding for water-saving and drought-resistance rice (WDR) in China. *Journal of Experimental Botany*, 61(13), 3509–3517.
- Manavalan, L. P., Guttikonda, S. K., Phan Tran, L. S., & Nguyen, H. T. (2009). Physiological and molecular approaches to improve drought resistance in soybean. *Plant & Cell Physiology*, 50(7), 1260–1276.
- May, L. H., & Milthorpe, F. L. (1962). Drought resistance of crop plants. *Field Crop*, 15, 171–179.
- Mehdi, S. S., & Ahsan, M. (2000). Genetic coefficient of variation, relative expected genetic advance and inter-relationships in maize (*Zea mays* L.) for green fodder purposes at seedling stage. *Pakistan Journal of Biological Science*, 3(11), 1890–1891.
- Min, H., Chen, C., Wei, S., Shang, X., Sun, M., Xia, R., ... Xie, Q. (2016). Identification of drought tolerant mechanisms in maize seedlings based on transcriptome analysis of recombination inbred lines. *Frontiers in Plant Science*, 7, 1080.
- Mittal, S., Mallikarjuna, M. G., Rao, A. R., Jain, P. A., Dash, P. K., & Thirunavukkarasu, N. (2017). Comparative analysis of CDPK family in maize, Arabidopsis, rice, and sorghum revealed potential targets for drought tolerance improvement. *Frontiers in Chemistry*, 5, 115.
- Muraya, M. M., Ndirangu, C. M., & Omolo, E. O. (2006). Heterosis and combining ability in diallel crosses involving maize (*Zea mays* L.) S1 lines. *Australian Journal of Experimental Agriculture*, 46, 387–394.
- Nielsen, R. L. (2007). *Assessing effects of drought on corn grain yield*. West Lafayette, IN: Purdue University.
- Obidiegwu, J. E., Bryan, G. J., Jones, H. G., & Prashar, A. (2015). Coping with drought: stress and adaptive responses in potato and perspectives for improvement. *Frontiers in Plant Science*, 6, 542.
- Pantuwan, G., Fukai, S., Cooper, M., Rajatasereekul, S., & O Toole, J. C. (2002). Yield response of rice (*Oryza sativa* L.) genotypes to different types of drought under rainfed lowlands: Part 1. Grain yield and yield components. *Field Crops Research*, 73(2–3), 153–168.
- Petrozza, A., Santaniello, A., Summerer, S., Di Tommaso, G., Di Tommaso, D., Paparelli, E., ... Cellini, F. (2014). Physiological responses to Megafol treatments in tomato plants under drought stress: A phenomic and molecular approach. *Scientia Horticulturae*, 174, 185–192.
- Phillippe, R., Courtois, B., McNally, K. L., Mournet, P., El-Malki, R., Le Paslier, M. C., & This, D. (2010). Structure, allelic diversity and selection of Asr genes, candidate for drought tolerance, in *Oryza sativa* L. and wild relatives. *Theoretical and Applied Genetics*, 121(4), 769–787.
- Pingali, P. L., & Heisey, P. W. (2001). Cereal-crop productivity in developing countries: Past trends and future prospects. *Agricultural Science Policy: Changing Global Agendas*, 99–03.

- Poroyko, V., Hejlek, L. G., Spollen, W. G., Springer, G. K., Nguyen, H. T., Sharp, R. E., & Bohnert, H. J. (2005). The maize root transcriptome by serial analysis of gene expression. *Plant Physiology*, *138*(3), 1700–1710.
- Qayyum, A., Ahmad, S., Liaqat, S., Malik, W., Noor, E., Saeed, H. M., & Hanif, M. (2003). Screening for drought tolerance in maize (*Zea mays* L.) hybrids at an early seedling stage. *African Journal of Agricultural Research*, *7*, 3594–3604.
- Rahaman, M., Chen, D., Gillani, Z., Klukas, C., & Chen, M. (2015). Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Frontiers in Plant Science*, *6*, 619.
- Rensink, W. A., & Buell, C. R. (2005). Microarray expression profiling resources for plant genomics. *Trends in Plant Science*, *10*(12), 603–609.
- Richards R.A. (1985). *Physiology and the breeding of winter-grown cereals for dry areas*. In J.P. Srivastana, E. Porceddu, E. Acevedo and S. Varma (eds.). In *Proceedings of an international workshop on drought tolerance in winter cereals 1987*; 27–31 Oct., Capri, Italy, (pp. 171–190).
- Romano, G., Zia, S., Spreer, W., Sanchez, C., Cairns, J., Araus, J. L., & Müller, J. (2011). Use of thermography for high throughput phenotyping of tropical maize adaptation in water stress. *Computers and Electronics in Agriculture*, *79*(1), 67–74.
- Saif-ul-malook, Ali, Q., Shakeel, A., Sajjad, M., & Bashir, I. (2014). Genetic variability and correlation among various morphological traits in students of UAF, Punjab Pakistan. *International Journal of Case Reports*, *1*, 1–4.
- Saijo, Y., Hata, S., Kyojuka, J., Shimamoto, K., & Izui, K. (2000). Over-expression of a single Ca²⁺-dependent protein kinase confers both cold and salt/drought tolerance on rice plants. *The Plant Journal*, *23*(3), 319–327.
- Saini, H. S., & Westgate, M. E. (2000). Reproductive development in grain crops during drought. In D. L. Spartes (Ed.), *Advances in agronomy* (pp. 59–96). San Diego, CA, USA: Academic Press.
- Scarpeci, T. E., Frea, V. S., Zanor, M. I., & Valle, E. M. (2017). Overexpression of AtERF019 delays plant growth and senescence, and improves drought tolerance in Arabidopsis. *Journal of Experimental Botany*, *68*(3), 673–685.
- Schussler, J. R., & Westgate, M. E. (1991). Maize kernel set at low water potential: II. Sensitivity to reduced assimilates at pollination. *Crop Science*, *31*(5), 1196–1203.
- Sharma, S., & Carena, M. J. (2016). BRACE: A method for high throughput maize phenotyping of root traits for short-season drought tolerance. *Crop Science*, *56*(6), 2996–3004.
- Shaw R.H. (1983). Estimates of yield reductions in corn caused by water and temperature stress. *Crop reactions to water and temperature stresses in humid, temperate climates*; (pp. 49–66).
- Sheikh, F. A. (2017). Recent advances in QTL mapping and quantitative disease resistance approach. *International Journal of Current Microbiology and Applied Sciences*, *6*(4), 1967–1984.
- Shikha, M., Kanika, A., Rao, A. R., Mallikarjuna, M. G., Gupta, H. S., & Nepolean, T. (2017). Genomic selection for drought tolerance using genome-wide SNPs in maize. *Frontiers in Plant Science*, *8*, 550.
- Singh, V., van Oosterom, E. J., Jordan, D. R., & Hammer, G. L. (2012). Genetic control of nodal root angle in sorghum and its implications on water extraction. *European Journal of Agronomy*, *42*, 3–10.
- Song, K., Kim, H. C., Shin, S., Kim, K. H., Moon, J. C., Kim, J. Y., & Lee, B. M. (2017). Transcriptome analysis of flowering time genes under drought stress in maize leaves. *Frontiers in Plant Science*, *8*, 267.
- Soni, P., Nutan, K. K., Soda, N., Nongpiur, R. C., Roy, S., Singla-Pareek, S. L., & Pareek, A. (2015). *Towards understanding abiotic stress signaling in plants: Convergence of genomic, transcriptomic, proteomic, and metabolomic approaches. Elucidation of abiotic stress signaling in plants* (pp. 3–40). New York, NY: Springer.
- Statista [Internet]. (2020) Oct 30 [cited 2021]. <<https://www.statista.com/topics/986/corn/#dossierSummary>>.
- Tardieu, F. (2013). Plant response to environmental conditions: assessing potential production, water demand, and negative effects of water deficit. *Frontiers in Physiology*, *4*, 17.
- Thirunavukkarasu, N., Hossain, F., Arora, K., Sharma, R., Shiriga, K., Mittal, S., et al. (2014). Functional mechanisms of drought tolerance in subtropical maize (*Zea mays* L.) identified using genome-wide association mapping. *BMC Genomics*, *15*(1), 1–12.
- Torun, M., & Koycu, C. (1999). Study to determine the relationship between grain yield and certain yield components of maize using correlation and path coefficient analysis. *Turkish Journal of Agriculture and Forestry*, *23*, 1021–1027.
- Turner, N. C. (1979). Drought resistance and adaptation to water deficits in crop plants. *Stress Physiology Crop Plants*, 343–372.
- Umakanth, A. V., Satyanarayana, E., & Kumar, M. V. (2000). Correlation and heritability studies in Ashwini maize composite. *Annals of Agricultural Research*, *21*, 228–230.
- Upadhyay, N., Kar, D., Deepak Mahajan, B., Nanda, S., Rahiman, R., Panchakshari, N., et al. (2019). The multitasking abilities of MATE transporters in plants. *Journal of Experimental Botany*, *70*(18), 4643–4656.
- Vaezi, S., Mishani, A., Samadi, Y., & Ghannadhs, M. R. (2000). Correlation and path coefficient analysis of grain yield and its components. *Iranian Journal of Agriculture Science*, *31*, 71–83.
- Vankova, R., Dobra, J., & Storchova, H. (2012). Recovery from drought stress in tobacco: an active process associated with the reversal of senescence in some plant parts and the sacrifice of others. *Plant Signaling & Behavior*, *7*(1), 19–21.
- Wang, C. T., Ru, J. N., Liu, Y. W., Yang, J. F., Li, M., Xu, Z. S., & Fu, J. D. (2018). The maize WRKY transcription factor ZmWRKY40 confers drought resistance in transgenic Arabidopsis. *International Journal of Molecular Sciences*, *19*(9), 2580.
- Wang, Y., Duan, L., Lu, M., Li, Z., Wang, M., & Zhai, Z. (2006). Expression of NAC1 up-stream regulatory region and its relationship to the lateral root initiation induced by gibberellins and auxins. *Science China Life Sciences*, *49*(5), 429–435.
- Wang, Z., & Huang, B. (2004). Physiological recovery of Kentucky bluegrass from simultaneous drought and heat stress. *Crop Science*, *44*(5), 1729–1736.

- Waseem, M., Ali, Q., Ali, A., Samiullah, T. R., Ahmad, S., Baloch, D. M., & Bajwa, K. S. (2014). Genetic analysis for various traits of *Cicer arietinum* under different spacing. *Life Sciences*, *11*(12s), 14–21.
- Westgate M.E., Bassetti P. (1990). Heat and drought stress in corn: What really happens to the corn plant at pollination. In *Proceedings of the forty-fifth annual corn and sorghum research conference, Chicago, IL*, (pp. 12–28).
- Westgate, M. E., & Boyer, J. S. (1985). Carbohydrate reserves and reproductive development at low leaf water potentials in Maize 1. *Crop Science*, *25*(5), 762–769.
- Xia, Z., Liu, Q., Wu, J., & Ding, J. (2012). ZmRFP1, the putative ortholog of SDIR1, encodes a RING-H2 E3 ubiquitin ligase and responds to drought stress in an ABA-dependent manner in maize. *Gene*, *495*(2), 146–153.
- Xu, J., Tian, Y. S., Peng, R. H., Xiong, A. S., Zhu, B., Jin, X. F., et al. (2010). AtCPK6, a functionally redundant and positive regulator involved in salt/drought stress tolerance in Arabidopsis. *Planta*, *231*(6), 1251–1260.
- Yousafzai, F., Al-Kaff, N., & Moore, G. (2009). The molecular features of chromosome pairing at meiosis: The polyploidy challenge using wheat as a reference. *Functional & Integrative Genomics*, *10*, 147–156.
- Yue, B., Xue, W., Xiong, L., Yu, X., Luo, L., Cui, K., & Zhang, Q. (2006). Genetic basis of drought resistance at reproductive stage in rice: Separation of drought tolerance from drought avoidance. *Genetics*, *172*(2), 1213–1228.
- Zenda, T., Liu, S., Wang, X., Liu, G., Jin, H., Dong, A., . . . Duan, H. (2019). Key maize drought-responsive genes and pathways revealed by comparative transcriptome and physiological analyses of contrasting inbred lines. *International Journal of Molecular Sciences*, *20*(6), 1268.
- Zhan, A., Schneider, H., & Lynch, J. P. (2015). Reduced lateral root branching density improves drought tolerance in maize. *Plant Physiology*, *168*(4), 1603–1615.
- Zhang, Q., Liu, H., Wu, X., & Wang, W. (2020). Identification of drought tolerant mechanisms in a drought-tolerant maize mutant based on physiological, biochemical and transcriptomic analyses. *BMC Plant Biology*, *20*(1), 1–14.
- Zhou, Q., Dong, Y., Shi, Q., et al. (2017). Verification and fine mapping of qGW1.05, a major QTL for grain weight in maize (*Zea mays* L.). *Molecular Genetics and Genomics*, *292*, 871–881.
- Zhou, X. H., Cheng, Y. X., Yaohal, Y., & Young, G. Z. (2004). Study on heterosis utilization of maize inbred lines in different ecological areas. *Journal of Maize Sciences*, *12*(4), 35–38.
- Zinselmeier, C., Habben, J. E., Westgate, M. E., & Boyer, J. S. (2000). Carbohydrate metabolism in setting and aborting maize ovaries. In M. Westgate, & K. Boote (Eds.), *Physiology and modeling kernel set in maize* (pp. 1–13). Madison, USA: CSSA and ASA.

Deciphering the genomic hotspots in wheat for key breeding traits using comparative and structural genomics

Dharmendra Singh¹, Pritesh Vyas², Chandranandani Negi², Imran Sheikh² and Kunal Mukhopadhyay³

¹Government Model College, Jhabua, Madhya Pradesh, India, ²Dr. Khem Singh Gill Akal College of Agriculture, Eternal University, Baru Sahib, Himachal Pradesh, India, ³Department of Bioengineering and Biotechnology, Birla Institute of Technology, Mesra, Jharkhand, India

19.1 Introduction

Wheat (*Triticum aestivum* L., $2n = 6x = 42$, AABBDD) is a dietary staple of 35% of the world's population and provides ~20% of the protein consumed by humans (Shiferaw et al., 2013). The bread wheat constitutes about 95% of the globally cultivated wheat and the remainder 5% is durum wheat, which is mostly grown in the Mediterranean region (Shewry, 2009). Bread wheat has a large and complex allopolyploid genome of 17 Gb size, having >80% repetitive and 20% structural and functional sequences. The intrachromosomal duplications of 24% of the total genes further enhance the complexity of the genome (Uauy, 2017). In the last years the primary focus of wheat breeding was improving yield, end-use quality, and resistance to certain stresses. However, continuous past selection for limited number of traits led to narrowness in the genetic base of wheat, making it vulnerable to various stresses (i.e., biotic and abiotic) (Mujeeb-Kazi et al., 2017). The rate of increase in wheat yield has been about 0.9% per annum, which is far less than the required 2.4% increase needed to feed more than 9 billion humans by 2050 (Ray, Mueller, West, & Foley, 2013). The challenge of meeting the target becomes even more difficult under climate change events, scarcity of water, and shrinkage of arable land (Daryanto, Wang, & Jacinthe, 2016). Recently, the crop modeling studies have predicted yield reductions of 6%–13%, with a 1°C rise in temperatures. To tackle the concerns associated with climate change and to develop abiotic/biotic stress-resilient cultivars, the genetic base of cultivated wheat needs to be urgently broadened (Ceoloni et al., 2017). This could be achieved by identifying the resistant and tolerant genes or quantitative trait loci (QTLs) using the advanced sequencing technologies and bioinformatics tools and transferring them to elite wheat cultivars using wheat prebreeding programs. The location of such genes or QTLs on wheat chromosomes are considered to be hotspots for the respective trait. Most of the tools or pipelines used in analysis are objective oriented and take advantage of model crop system for predicting genes using comparative genomics.

The comparative genomics is an approach to compare the complete genome sequences of different species using different alignment tools. Identifying “conserved” DNA sequences is an important step toward understanding the genome itself. It pinpoints genes that are essential to life and highlights genomic signals that control gene function across many species. Additionally, it helps us to further understand what genes relate to various biological systems, which in turn may translate into novel mechanism of stress tolerance in plants (<https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet>). The comparative genomics in wheat covers the study of evolution and isolation or characterization of genes using the rice genome (Gupta, Pandey, Gopalareddy, Sharma, & Singh, 2019). Recent advancements in experimental approaches, resources, and computational analysis tools have facilitated the identification of new genes that can be utilized in wheat breeding. In this chapter, we focused on genomic comparisons, functional comparative genomics, gene discovery, and marker developments in wheat using rice as model system. Furthermore, we highlighted the genomic hotspots in wheat considering the adaptive and agronomic traits. Finally, we discussed how genomic hotspots were identified, started from genomic sequences using different methods and tools. Identification of genomic hotspots will significantly assist in wheat improvement programs.

19.2 Genomic comparisons and gene discovery

The grass family includes cereals such as wheat, barley, maize, sorghum, millet, and rice that are the most important crops for human and animal nutrition. Over the last decades, significant findings reported for cereal comparative genomic and it pioneered the field of plant comparative genomics. First, the comparative studies were conducted at the genetic map stage and shown a very strong collinearity of molecular markers and QTL along the chromosomes associated with agronomic characteristics, thus establishing evolutionary relationships between the cereal genomes. Rice among cereals was the first to be selected for genome sequencing due to its small size and was used as a model crop for such studies. The first comparative study of intragenomic relationships has shown several micro-colinearity disrupting rearrangements in the past 50–70 million years that is attributed to improvements in the development of large bacterial artificial chromosome (BAC) arrays and BAC sequences.

In the last few decades, mapping and defining clusters with identical gene orders in cereals has provided strong evidence of gene order persistence across several megabases, referred to as macro-colinearity. The macro-colinearity across different genomes is summarized and represented through the “Circle Diagram” and the analysis is termed synteny analysis (Gale & Devos, 1998). The synteny map for wheat, rice, oats, maize, sorghum, sugarcane, foxtail millet, and finger millet was later expanded to include 10 grass species using less than 30 rice linkage groups (Devos & Gale, 2000). The study of genomic variations and evolutionary divergence across 60 million years are noteworthy and reflected as the variations in the size of grass genomes such as wheat (17,000 Mb), rice (430 Mb), sorghum (770 Mb), and maize (2700 Mb) (Arumuganathan & Earle, 1991; Gale & Devos, 1998; Keller & Feuillet, 2000). The initial work on colinearity of genetic markers was greatly improved after the identification of colinearity pattern among multiple plant genomes for agronomically important trait-linked genetic markers (Peng et al., 1990). Several recent cereal studies have reported incomplete micro-colinearity at the sequence level (SanMiguel, Ramakrishna, Bennetzen, Busso, & Dubcovsky, 2002; Tikhonov, Bennetzen, & Avramova, 2000). Song, Llaca, and Messing (2002) described the comprehensive micro-colinearity as orthologous regions in maize, sorghum, and the two rice subspecies. The study indicated that the gross macro-colinearity is preserved, but micro-colinearity is incomplete among these cereals. Micro-rearrangement or small-scale genomic shifts, such as gene insertions, deletions, duplications, or inversions, are due to the deviations from gene colinearity (Bancroft, 2000). The synteny study revealed that 6 genes in rice, 15 genes in sorghum, and 13 genes in maize were present in the orthologous region (Song et al., 2002). Gene amplification triggered a local expansion of conserved genes in maize and sorghum but did not interrupt their order or orientation. As predicted, the two shotgun-sequenced rice subspecies, japonica, and indica, which diverged over 1 million years ago, have a high degree of gene conservation between them (Bennetzen, 2000). However, narrow regions of divergence can be detected in these genomes upon careful examination (Song et al., 2002). These regions correspond to the areas of increased differentiation among rice, sorghum, and maize, implying that it may be useful to align the two rice subspecies to distinguish regions of cereal genomes vulnerable to rapid evolution. The discrepancies between the genes of sorghum and maize emerged after the two species’ ancestral genomes diverged from each other 16.5 million years ago (Gaut & Doebley, 1997). The region where micro-colinearity is broken: a gene is “missing” from its orthologous position, however a matching gene homolog may sometimes be found in a nonorthologous place (Xu, Lagudah, Moose, & Riechers, 2002). In wheat, two copies of a gene are duplicated to give rise to two separate glutathione transferase genes. In the orthologous wheat site, the best-conserved copy of the rice gene was not located. Alternatively, it was located on rice chromosome 10 at nonorthologous location (Xu et al., 2002). Song et al. (2002) showed that gene amplification and gene translocation are associated functionally, and that these differential genome divergences were manifested during speciation. When orthologous areas of different species are compared, a mosaic of conserved segments interspersed with nonconserved segments becomes evident (Song et al., 2002). The information obtained through comparative genome studies in terms of the orthologous regions has been used to select candidate genes associated with agronomic traits and to select molecular markers to increase the density of the map at particular genetic locations, thus enabling map-based cloning.

19.2.1 Gene discovery and marker development

Comparative genomic studies have contributed to the development of genomic instruments that can be used to understand genome structure and genetic architecture, tailoring the productive genetic studies and techniques for gene isolation.

19.2.1.1 Colinearity-based gene cloning

In some cases, the retention of the gene sequence at orthologous locations reflects conservation of a common gene feature between the species. Early comparative genomics studies found a variety of genes at orthologous locations in cereal

genomes responsible for developmental and domestication traits, such as shattering, plant height, vernalization, flowering time, row number, and kernels per row (Bailey, West, & Black, 2015; Paterson et al., 1995). Researchers have isolated the rice genes and their sequence information was used in positional cloning of the orthologous genes in other genomes (Kilian, Chen, Han, Steffenson, & Kleinjans, 1997). The isolation of the “green revolution” dwarfing genes *Sd1* (Monna et al., 2002), *Rht-1* in wheat, *D8* in maize is the example of application of colinearity in gene form and function using rice sequence information for gene cloning in other cereals (Peng et al., 1990). The isolation of genes in barley, wheat, and maize has also revealed examples of genes retained at orthologous locations in cereals. The examples are wheat vernalization gene, *Vrn1* (Yan et al., 2003), and the barley photoperiod *PPD-H1* gene (Turner, Beales, Faure, Dunford, & Laurie, 2005) retained in orthologous region in rice. In addition, for similar phenotypes, gene conservation concept was proposed based on the protection of genetic positions, for example, the mutation of maize bare stalk1 was mapped in a colinear region with the rice lax panicle gene (Gallavotti et al., 2004). Candidate genes can be detected directly from the rice sequence in regions where micro-colinearity is high, even if the target trait has not been mapped to the colinear location in rice. This was successfully used to help the isolation of the *Ror2* (Collins, Thordal-Christensen, Lipka, & Bau, 2003) gene for powdery mildew resistance and the *sw3* dwarfism gene in barley (Gottwald, Stein, Borner, Sasaki, & Graner, 2004). Furthermore, similar roles did not appear to be linked with similar genes have shown in a study by Griffiths, Dunford, Coupland, and Laurie (2003), comparing QTL for heading time in rice and barley. In rice, several flowering QTLs belong to the *CONSTANS* gene family but none of the homologous *CONSTANS* genes are associated with any of the known QTL for flowering time in barley. In general, the information of colinearity and conserved regions comprising characterized genes and QTLs involved in developmental processes and selected during domestication in cereal genomes are good candidates for direct gene isolation based on the comparative genomics. The genes of cereals do not show colinearity between the genomes of cereals and evident from genes in grasses for disease resistance, where map-based cloning of rice genes has not been benefited much from the grass genes knowledge. By comparative genetic analysis the nonsyntenic position of these genes between cereals has already been established (Leister et al., 1998), and in several instances, attempts to use colinearity with rice to isolate R genes have shown the limits of colinearity between cereal genomes. The first case that questioned the use of rice colinearity for map-based cloning of R genes was working with the barley stem rust resistance gene *Rpg1*. No orthologous genes were found throughout the rice genome; nonetheless, a certain degree of colinearity was retained in the orthologous locus in rice (Kilian et al., 1997), and *Rpg1* map-based cloning in barley has been achieved (Brueggeman et al., 2002). In certain cases, the nonorthologous locations of homologous genes of wheat and rice suggest major genome rearrangements, such as with the leaf rust *Lr10* and the powdery mildew *Pm3* fungal disease R genes (Guyot, Yahiaoui, Feuillet, & Keller, 2004). Both these genes have been cloned using alternate methods. Chen et al. (2005) have recently documented colinearity between a QTL that confers resistance to the blast fungus, *Magnaporthe grisea*, and its homologous role on barley and rice. Regardless of whether rice possesses the gene at its orthologous position, it will be necessary to saturate the gene’s target area with flanking regions in other cereal genomes. Rice expressed sequence tags (ESTs), for example, were used to reduce the genetic interval around the R loci *Rpg1* and *Rph7* so that chromosome walking could proceed in barley (Brunner, Keller, & Feuillet, 2003). Using a map-based cloning technique, markers are used to construct a genetic map for Bru1, which is then used to cross-pollinate with other grains to create resistance (Asnaghi et al., 2004). There are some more examples of the use of EST rice-related markers to saturate genetic areas in other cereals, and this approach is now widely used in laboratories engaged in the cloning of global cereal genes (Bortiri, Jackson, & Hake, 2006; Collins et al., 2003; Yan et al., 2004). After rice, a new model species, *Brachypodium*, has recently been proposed for temperate cereals such as wheat and barley (Draper et al., 2001; Vogel et al., 2006). *Brachypodium* has been used successfully in conjunction with rice to isolate *Ph1*, one of the primary genes regulating pairing of chromosomes in polyploid wheat (Griffiths et al., 2006). Different techniques, such as using closely related plants, or mapping particular genes in species of interest, are applied under very low colinearity. Such an example is provided by the maize *Ramosal* gene that controls the architecture of the tassel. This gene is unique to the tribe of the *Andropogoneae* and lacks in rice which was recently isolated using a transposon-tagging method (Vollbrecht, Springer, Goh, Buckler, & Martienssen, 2005). The so-called subgenome map-based cloning in wheat (Stein, Feuillet, Wicker, Schlegelhauf, & Keller, 2000) was used to isolate the *Lr10* and *Pm3* disease R genes, both of which are located in a noncolinear rice region on the short arm of chromosome 1A (Guyot et al., 2004). Genetic mapping was conducted on hexaploid wheat using chromosome walking with BACs from a diploid relative of wheat, *Triticum monococcum* (Yahiaoui, Srichumpa, Dudler, & Keller, 2004).

19.2.2 Gene annotation and marker development

Understanding gene structure and its role within an organism is dependent on the whole-genome sequencing. Since genes are the most conserved features of genomes, reference-annotated genomes from other species can be used to

predict the query genome. This is evident from comparison between distant genomes, such as rice and *Arabidopsis thaliana*, whose ancestors diverged 200 million years ago, a significant number of genes have been retained (Salse, Piegu, Cooke, & Delseny, 2002). These findings support the use of rice genome sequence as annotated reference for characterization and annotation of cereal genomes. Further, aligning other cereal species ESTs with the rice genome can better predict new rice genes. Generating a large collection of full-length cDNAs in rice can be used to confirm intron/exon boundaries and can be used to train gene predictors. Creation of new markers for intron/exon boundaries is also useful. Indeed, the polymerase chain reaction (PCR) frequency is higher in introns than in exons, and it is simpler to design PCR primers that amplify intron sequences in organisms that chronically suffer from a lack of polymorphism. This principle has recently been introduced in pearl millet by Bertin, Zhu, and Gale (2005) and reported association of millet EST sequences with rice transcript sequences, to predict the position of introns which were then amplified followed by the discovery of single-strand conformation polymorphism (SSCP). The technique of the SSCP-SNP marker has considerable potential for the development of COS (Conserved Orthologous Set) markers for comparative cereal mapping.

19.2.3 Functional comparative genomics in cereals

The development of genomic tools, in particular EST collections for several crops, has enabled the study of gene expression based on the DNA chips. DNA chips are now available for major cereals such as rice, wheat, barley, and maize. The comparisons of DNA chip experiments were initially difficult due to variations in sampling and conditions. However, technical advancements in robotics and development of high-quality DNA, standardization, and normalization process made the experimental comparisons possible (Brazma & Vilo, 2001), while some limitations remain, namely, inadequate annotation, incomplete representation of the genome for most crops, different developmental kinetics and phenotypic stages, variable experimental conditions. However, comparison of gene expression profile is possible in different cereals under similar physiological and biological circumstances. Several stress response transcriptomic experiments using various types of DNA chips have been conducted in different cereals, but results were not compared extensively. Comparative microarray experiments have led to the discovery of 65 candidate genes differentially expressed in winter wheat under cold stress (Gulick et al., 2005). Expression profiling studies in cereals (<http://barley-base.org/>; <http://www.ricearray.org/>; <http://www.maizearray.org/>) can accumulate and be processed in databases, allowing the rapid creation of metaanalysis of expression trends throughout cereal organisms.

19.3 Genomic hotspots in wheat

19.3.1 Biofortification hotspots

After rice and maize, wheat is an essential crop utilized by people and dominates the cereal production globally. To live a disease-free and productive life, a human being needs around 44 nutrients in sufficient quantity, which are obtained from a well-balanced nutritious diet. In developing countries, people are majorly dependent on plant-driven food, which is low in micronutrient content, resulting in around 2 billion population suffering from malnutrition globally (Sheikh, Vyas, & Dhaliwal, 2020). In this approximately 60% of world population suffers from iron while 30% suffers from zinc deficiency (White & Broadley, 2009). Since these people cannot afford quality food rich in vitamins and minerals, they suffer from micronutrient deficiencies, and make it the fifth major global challenge for human health as listed in Copenhagen Consensus. Iron deficiency leads to anemia and disruption of proper functioning of immune and endocrine system, with prime targets being children and pregnant women. Insufficient dietary zinc intake can be noticed in the form of retarded growth, development, unexplainable weight loss, and depression. Since zinc does not have long-term storage in body, its regular dietary intake is necessary. To counter these problems, global community has set Sustainable Development Goals aiming to end the nutritional deficiencies in all of its form. Also, this zero hunger goal aims to end the hunger with better food in terms of nutrition.

Biofortification refers to the strategies involved for increasing the levels of vitamins and minerals in edible parts of the crops using various techniques of biotechnology, genomics and breeding. It is a long-term approach for removing micronutrients deficiency. Since wheat is used as food by billions of people, increasing grain Fe and Zn would significantly impact the health by easing such deficiencies (Ali & Borrill, 2020). Also, because iron and zinc are abundantly found minerals in humans and are present in similar dietary sources, they are evaluated together. Biofortification has nonrecurrent expense with sustainable way to combat starvation-related issues. Augmentation of mineral concentration through the use of foliar sprays is mostly done but this adds to the high expenditure of farmers. Utilization of breeding methods to

achieve genetic variation is readily attainable through primary, secondary, and tertiary gene pool of the crop. However, genetic transformation is better approach when genetic variation is absent or unattainable to exploit. Current achievements in genomic biofortification gave an enormous potential to provide sustainable solutions for issues relating to hidden hunger. These genomic strategies include genomic selections, marker-assisted selection, and QTL mapping.

Already available genetic variations among different species and landraces are used for the biofortification breeding program for developing nutrient-enriched wheat germplasm along with higher yield potential and stress resistance. Although grain such as Fe content (GFeC) and zinc content (GZnC) is an essential objective in breeding program, traits like grain yield and grain protein content (GPC) should not be traded off because the farmer income is completely dependent on the yield. In the view of major climatic changes prevailing and being predicted in the future, how the yield and GPC would be influenced by these changes would be an utmost priority. In consideration to these changes, improving yield potential and developing stress-tolerant wheat is a paramount task. Utilization of next-generation sequencing, marker-assisted selection with advanced molecular techniques has been a promising approach to attain nutrient-rich varieties.

In the year 1997, mapping of first QTL for iron and zinc in wheat was done, on chromosome 6BS in the population produced from cross between durum wheat and wild emmer (Joppa, Du, Hart, & Hareland, 1997). This QTL, Gpc-B1, imparted 18% and 12% increment in Fe and Zn, respectively. The gene underlying the QTL was found to be nascent polypeptide-associated complex (NAC) transcription factor-B1 (Distelfeld et al., 2007; Uauy, Distelfeld, Fahima, Blechl, & Dubcovsky, 2006). QTLs for resistance to abiotic and biotic stresses have been positively influenced by the development of powerful technologies for genome sequencing and genome-wide DNA markers. Identifying necessary QTLs, with specific introgression into varieties is a faster approach for development of abiotic or biotic stress-resistant versions of mostly used susceptible varieties. Till date, four zinc biofortified varieties are released, Zinc Shakti, which is developed through transfer of genes from *Aegilops squarrosa* into PBW343, Zincol-2016, developed by transfer of *Triticum spelta* genes into NARC2011, and HPBW-010 and WB020 developed through transfer of genes from *A. squarrosa* and *Triticum dicoccon* (Singh et al., 2017). Identification of QTLs for Fe and Zn has been done in various works (Genc et al., 2009; Shi et al., 2008; Tiwari et al., 2009, 2016; Yan et al., 2018). However, linkage drag of low yield and harvest index is the major challenge in commercial use. Velu et al. (2018) recognized QTL on 2 and 7 group chromosomes for a population of 330 spring common wheat varieties. Alomari et al. (2018) identified markers related to higher zinc concentration across majority of chromosomes but marker–trait associations (MTAs) were most significant on chromosomes 5A and 3B. Genes involved in Zn transportation and for basic leucine zipper (bZIP) and mitogen-activated protein kinase (MAPK) were also present in these genomic areas. Various studies have been conducted in wheat to identify QTL for micronutrients as well as macronutrients like Ca (Alomari, Eggert, Von Wirén, Pillen, & Röder, 2017; Crespo-Herrera, Velu, & Singh, 2016; Morgounov et al., 2007; Tiwari et al., 2009; Xu et al., 2012b). Krishnappa et al. (2017) investigated genomic region each on 5A (Xgwm126-Xgwm595) and 7A (Xbarc49-Xwmc525) containing QTL for both Fe and Zn. Peleg et al. (2009) identified 6 QTLs on 2A, 2B, 3A, 4B, 5A, 6A, 6B, 7A, and 7B chromosomes for micronutrients (zinc, iron, copper, and manganese) and macronutrients (calcium, magnesium, potassium, phosphorous, and sulfur) in durum wheat x emmer wheat recombinant inbred lines population. Genc et al. (2009) recognized four QTLs for GZnC on chromosomes 3D, 4B, 6B, and 7A in a doubled haploid wheat population. Similar work has been done for Fe, Zn (Hao, Velu, Peña, Singh, & Singh, 2014; Srinivasa et al., 2014; Xu et al., 2012b). However, low resolution of QTL obtained using biparental populations and dependent on the genomic variation in two parents is used for deriving the mapping population. Sheikh et al. (2018) identified the metal homeostasis genes located on chromosomes of the homologous groups 2 and 7 in wheat using intron-targeted amplified polymorphism markers. These genes control root acquisition, accumulation, and movement of micronutrients in the crops.

Uptake and translocation of zinc is affected by activity of transporter proteins which are present in plasma membrane of root cells. These transporter families include zinc-regulated and iron-regulated transporter-like proteins (ZIP), cation diffusion facilitator proteins, natural resistance-associated macrophage proteins, yellow stripe like, adenosine triphosphate (ATP)-binding cassette transporters, and heavy-metal ATPases (Xia et al., 2020). Elevated expression of zinc transporter TdZIP1 was found in wild emmer wheat under zinc deficiency (Durmaz et al., 2011). Evens, Buchner, Williams, and Hawkesford (2017) revealed that *T. aestivum* consists of various group FbZIP transcription factors which alter the expression of ZIPs through binding to Zn-deficiency response elements in their promoter. Exploring the expression pattern of such genes would expand our knowledge about complex homeotic connections. Also, the concentrations of Fe and Zn are regulated through multiple genes, and their presence in the grain becomes a complex polygenic phenomenon (Kumar et al., 2018).

GPC being an essential trait contributing to the nutritional value and quality of end products of wheat makes it economically very important. Among all GPC QTLs studied, Gpc-B1 is the most significant, which after cloning studies confirmed the improvement of protein, Fe, and Zn concentrations by 38%, 18%, and 12%, respectively (Distelfeld et al., 2007; Uauy et al., 2006). Introgressing GPC-B1 allele in the background of elite lines has led to various beneficial varieties in several countries (Tabbitta, Pearce, & Barneix, 2017). Efforts for enhancement in GPC through conventional breeding do not provide the desired results because of high impact of environmental factors on GPC, negative correlation between grain yield and GPC, the quantitative genetic control of GPC, and the low heritability (Balyan et al., 2013). The use of the effective genetic tools and statistical methods has led to identification and mapping of QTLs involved in epistatic interaction (Conti et al., 2011; Kulwal et al., 2005; Xu et al., 2012b; Zhao et al., 2010).

In cereals, phosphorus is stored in the form of phytic acid (PA) which accounts for nearly 70%–80% of total phosphorus within grain. PAs have a characteristic to form insoluble complexes with multicharge metal ions and consuming food with high PA causes problem in absorbing minerals present in the diet. Therefore breeding focus is directed toward reducing the PA content, to boost bioavailability of minerals. The PA accounts for an antinutritional property, thus breeders aimed at finding hotspots for lower PA and higher phytase levels. ICAR-IIWBR Karnal evaluated 400 genotypes for phytase in wheat grains and reported Indian varieties with 3.4-fold variation, while the synthetic hexaploid wheat with 5.9-fold variation in phytase level (Ram, Verma, & Sharma, 2010). In a similar study, variability in PA levels was also found, which involved 257 wheat genotypes, including 168 synthetic hexaploids and 89 wheat cultivars (Vashishth, Ram, & Beniwal, 2017). Similarly, biofortification of wheat for anthocyanin is also a center of attention. The presence of anthocyanin greatly benefits the consumer because of its antioxidant properties and role in prevention of cardiovascular problems, cancer, obesity, and diabetes. Garg et al. (2016) developed such colored wheat lines, that is, black, purple, and blue, having high concentration of phenolics with considerable yield potential.

All such traits are influenced by numerous genetic as well as environmental factors. The use of genome-wide association (GWA) mapping helps in identifying associations between genotypes and phenotypes by utilizing unrelated individuals which have been phenotyped and genotyped simultaneously (Hirschhorn & Daly, 2005). Also, genome-wide association study (GWAS) led to improve QTL resolution by considering more representative and varied gene pool. In fact, this study permits the identification of nonrandom associations between genotype and phenotype in group of individuals with detection of genetic variants associated with compound agricultural traits. Work involving GWAS for Fe, Zn, carotenoid content, and GPC is also available in wheat (Gahlaut, Jaiswal, Singh, Balyan, & Gupta, 2019; Kumar et al., 2018). Bhatta, Morgounov, Belamkar, Yorgancilar, and Baenziger (2018) conducted GWAS for 10 grain minerals using synthetic hexaploid wheat, which consisted of 3 MTAs for Fe concentration on 1A and 3A while 13 MTAs for GZn concentration on 8 different chromosomes, 1A, 2A, 3A, 3B, 4A, 4B, 5A, and 6B. During GWAS, epistasis is usually ignored; however, in some studies, epistasis is studied (Jaiswal et al., 2016; Sehgal et al., 2017). Recently, multilocus and multitrait mixed model approaches are also being used to overcome this restrain. Also, when desired mutants are unavailable, gene editing through clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR associated protein 9 (Cas9) technology is efficiently used for the enhancement of yield with resistance toward abiotic and biotic stresses. CRISPR–Cas9 has been utilized in wheat for gene editing of TaDEP1, TaGW2, and TaGARS7 (Liang et al., 2017; Wang et al., 2018; Zhang et al., 2018).

19.3.2 Genomic hotspots for biotic stress resistance

Among various wheat diseases, leaf rust, stem rust, powdery mildew, yellow rust, and spot blotch are the most harmful. It is estimated that globally wheat rust pathogens cause economic loss of US\$ 4.3–5.0 billion (Figueroa, Hammond-Kosack, & Solomon, 2018). Rust pathogens severely affect the wheat yield and grain quality worldwide. The new races of rust pathogen keep evolving and spreading across the wheat-growing regions; therefore identification and use of the new sources of resistance genes are essential. Further, the genetic control is the most effective, economical, and environmentally safe method to minimize yield losses and controlling the disease. It is requisite to identify novel genes and pyramiding genes for different types of resistance to achieve high levels of durable resistance for sustainable control of the disease. There are three rust diseases of wheat, stem, leaf rust, and stripe rust all caused by members of the *Basidiomycete* family, genus *Puccinia*, named *P. graminis* f. sp. *tritici* (*Pgt*), *P. triticina* (*Pt*), *P. striiformis* f. sp. *tritici* (*Pst*), respectively (McIntosh, Wellings, & Park, 1995).

Leaf rust caused by *Puccinia triticina* majorly infects the foliar tissues, with circular to oval reddish-brown urediniospores or pustules produced on the infected leaves of wheat, favored by temperatures of 60°F–70°F. Genetic studies of leaf rust (*Lr*) resistance in wheat have been done by wheat researchers throughout the world. Mains, Leighty, and Johnston (1926) initially identified that the wheat cultivars Webster and Malakof had a gene which conditioned leaf rust resistance that was later designated as *Lr1* and *Lr2*, respectively. Sears (1956) used ionizing radiation to induce chromosome breakage and transferring a gene conditioning resistance to leaf rust from *Aegilops umbellulata* to wheat. Leaf rust (*Lr*) resistance genes named *Lr1* to *Lr68* have been characterized in wheat species with *Lr1*, *Lr3*, *Lr10*, and *Lr20* being most frequently used in wheat cultivars globally (Dakouri, McCallum, Radovanovic, & Cloutier, 2013). So far, 78 *Lr* genes have been cataloged in wheat (Gao et al., 2019). Some *Lr* resistance genes that have been cloned and sequenced include *Lr1* (Cloutier et al., 2007), *Lr10* (Feuillet et al., 2003), *Lr21* (Huang et al., 2003), and *Lr34* (Krattinger et al., 2009). The use of *Lr34* gene is being done in wheat breeding for more than a century. Due to its durable resistivity and broad-spectrum effectiveness, *Lr34* is one of the most common and frequently studied disease resistance genes in wheat breeding (Krattinger et al., 2019).

For breeding of disease-resistant wheat, two classes of genes are utilized. First class is referred to as *R* (resistance) genes, which are pathogen race specific in their action and are effective at all the stages of plant growth, with mostly encoding immune receptors of nucleotide-binding site leucine-rich repeat (NLR) class. These immune receptors recognize pathogen effector proteins which are delivered into the host cell during infection. Another class of genes comprises nonrace-specific or adult plant resistance (APR) genes, imparting resistivity only to the adult plant. In contrast to *R* genes, since the resistance imparted by APR genes occurs at later stage of plant growth, this leads to substantial disease development and as such resistance provided is partial. APR genes have the capability to provide resistance against all the isolates of a pathogen species and sometimes even have functionality against multiple pathogens. Till date, wheat genes conferring resistance to leaf (*Lr1*, *Lr10*, and *Lr21*), yellow (*Yr10*), and stem (*Sr22*, *Sr33*, *Sr35*, *Sr45* and *Sr50*) rust have been cloned, all of which encodes for NLR receptor proteins (Periyannan, Milne, Figueroa, Lagudah, & Dodds, 2017). However, in contrast to *R* genes, some APR genes have shown high durability, like the best known APR gene in wheat is the *Sr2*, which is functional against many races of stem rust pathogen for nearly 100 years (Ellis, Lagudah, Spielmeier, & Dodds, 2014). *Sr2* is the first APR gene for stem rust which is genetically defined in rust atlas (McIntosh et al., 1995).

Recently many APR genes have been cloned, which have given an insight into the mechanism involved in nonrace-specific resistance. For example, stripe rust resistance gene *Yr36*, encoding chloroplast-localized protein is suggested to reduce the detoxification of reactive oxygen species through phosphorylation of thylakoid-associated ascorbate peroxidase, which results in the increased defense response against the pathogen (Gou et al., 2015). Some other genes which were identified for conferring durable APR against multiple fungal diseases include very important genes, namely, *Lr34* (*Yr18/Sr57/Pm38*), *Lr46* (*Yr29/Sr58/Pm39*), and *Lr67* (*Yr46/Sr55/Pm46*). Expression of these genes results in partial resistivity against all the races of pathogens causing leaf, strip, stem rust, and powdery mildew. *Lr34* was initially reported as the modifier of APR in the cultivar Frontana. *Lr34* and *Sr2* have imparted partial resistance over many years in large areas with high and extended disease pressure, hence proving their durability (Ellis et al., 2014). The *Lr67* resistance allele encodes for a protein which has lost its hexose transport functionality, leading to disturbances in the sugar balance of intracellular and extracellular spaces of the leaf. This eventually reduces the accessibility of nutrients inside the host which is provided to the biotrophic fungi (Moore et al., 2015). None of the APR genes alone provide sufficient level of protection under high pathogen pressure and more often their expression can be too slow in the field for adequate yield protection.

Another wheat disease known as spot blotch (SB), caused by *Cochliobolus sativus* (anamorph: *Bipolaris sorokiniana*), is a devastating disease occurring in warm and humid regions like Africa, Latin America, Eastern India, China, and Southeast Asia. However, it has also been infecting the wheat growing in the Northwestern Russia indicating the occurrence of fungal virulence even in the European environment. The infection becomes intense during the grain filling stage, leading to remarkably high yield loss and deteriorated grain quality. The pathogen generates symptom on leaves, sheath, and stem. However, under extreme cases, it infects the spikelets, which results in shriveled grains with black coloration at the embryo end of kernel (Kumar, Joshi, Kumar, Chand, & Röder, 2010). This fungi forms thick-walled conidia for surviving through harsh conditions and inoculum overwinters on soil, wheat seeds, weeds, and rice stubbles. Healthy host tissues are punctured by the fungi, with conidiophores germinating within 4 h of infection and hyphae from preinfected cells entering the nearby intact host cells in 24 h. The detection of SB becomes visible in the form of dark brown lesions on the leaves (Jamil, Ali, Ali, & Mujeeb-Kazi, 2020). Mujeeb-Kazi et al. (2007) in their study suggested *A. tauschii* as a resistant source against SB pathogen. They developed an intergeneric cross product “Mayoor,” having germplasm which had two sources of resistance pyramided from *A. tauschii* and *Thinopyrum curvifolium*, which imparted considerable resistance against *C. sativus*. Lillemo, Joshi, Prasad, Chand, and Singh (2013)

mapped SB resistance gene *Sb1*, which was colocalized with leaf rust (LR) resistance locus *Lr34* on chromosome 7DS. Until now, three formally designated spot blotch (*Sb*) resistance genes (*Sb1–Sb3*) have been reported (Kumar et al., 2015; Lillemo et al., 2013; Lu et al., 2016). Recently, Zhang et al. (2020) identified and mapped a *Sb* gene, assigned as *Sb4*, against this pathogen in wheat.

A specific gene referred to as *ToxA* encoding for a host-selective toxin (HST) functions as a necrotrophic effector with often being responsible for the virulence of the pathogen. While the other known as *Tsn1* is a sensitivity gene present in the host plant, the presence of which helps in causing *ToxA*-positive pathogenicity resulting in spot blotch of wheat. Hence, wheat plants with the absence of *Tsn1* are generally resistant to spot blotch. The *ToxA* gene was first isolated from *P. tritici-repentis* (Balance, Lamari, Kowatsch, & Bernier, 1998; Ciuffetti, Tuori, & Gaventa, 1997) and was later identified in *B. sorokiniana* (McDonald, Ahren, Simpfendorfer, Milgate, & Solomon, 2018). Recently, Navathe et al. (2020) investigated the interaction of *Tsn1-ToxA* in wheat and indicated that the absence of *Tsn1* facilitated the resistance against SB of wheat. Hence, selection of wheat genotypes having the absence of *Tsn1* allele would improve the resistance to SB.

Stem rust (SR) disease caused by *Puccinia graminis* f. sp. *tritici* (*Pgt*) is one of the major diseases of wheat with wide distribution around the world. Visible reddish-brown oblong pustules with frayed margins on stems and leaves are observed on susceptible plants. Usually found in the regions having warm and moist conditions, this stem rust epidemic has occurred throughout major wheat-producing areas. McFadden (1930) during the devastating 1919 US stem rust epidemic developed a wheat variety known as Hope. This was released in the year 1926 and was derived from a cross between the SR susceptible North American cultivar Marquis and the highly SR-resistant tetraploid emmer wheat Yaroslav. Hope was identified for its quality and high-level SR resistance in field conditions. He, however, did not identify the *Sr2* gene specifically but selected a small combination of genes for stem rust resistance. Out of many stem rust (*Sr*) resistance genes, *Sr31* is the most extensively employed race-specific gene against *Pgt*. But eventually the evolution of pathogen to *Sr31* has led to the considerable use of other genes such as *Sr2*, *Sr25*, *Sr23*, *Sr33*, *Sr35*, *Sr45*, and *Sr50* (Singh, Govindan, & Andersson, 2017). Alien genes such as *Sr32* and *Sr39*, both derived from the short arm of chromosome 2 from different accessions of *Aegilops speltoides*, also serve resistance against many tested strains of the stem rust pathogen.

Stripe or yellow rust (YR) caused by *Puccinia striiformis* f. sp. *tritici* (*Pst*) is a destructive disease of wheat which has been reported in more than 60 countries (Chen, 2005). Presently, 88% of the wheat production in the world is susceptible to the stripe rust which leads to global loss of over 5 million tons of wheat which have an estimated market value of USD 1 billion dollars annually (Schwessinger, 2017). Stripe rust produces yellow linear pustules which run parallel with the leaf veins and the temperature of 50°F–60°F is favorable for the pathogen. To date, more than 70 stripe rust resistance genes, designated as *Yr* have been reported in wheat (Chen, 2005; Cheng & Chen, 2010). Initial approach to transfer small, nonhomologous alien segments for resistance was achieved by Riley (Riley, Chapman, & Johnson, 1968a,b). Riley and coworkers used a high pairing line of *A. speltoides* for induction of homoeologous recombination to transfer *Yr8*, a gene responsible for YR resistance, from *Aegilops comosa* ssp. *comosa* to wheat. Singh, Nelson, and Sorrells (2000) in their work identified and mapped a new gene from *A. tauschii*, designated as *Yr28*, which was involved in the seedling and field resistance to the predominant race of YR in the Mexican highlands. Recently, Li, Dundas, et al. (2020) observed a new yellow rust resistance gene *Yr83* on the rye chromosome in 6R in wheat. This gene showed a high level of seedling resistance to Australian pathotypes of the *Pst* pathogen and an even higher resistance to the Chinese *Pst* pathotypes in the field.

The Powdery mildew (PM) disease caused by biotrophic fungi *Erysiphe graminis* f. sp. *tritici* is a pathogen infecting many plant species, including important crops such as wheat, barley as well as the model plant *A. thaliana*. Since the suitable habitat conditions of stripe rust and powdery mildew are similar, there is a high probability of both diseases occurring in field simultaneously. Miller, Reader, Ainsworth, and Summers (1998) transferred gene *Pm12*, conferring resistance to powdery mildew from *A. speltoides* to wheat chromosome 6A; however, it merely contributed to cultivar improvement. Wild relatives of wheat constitute of huge pool of genes imparting desirable traits which can be exploited for wheat improvement. For instance, many wild relatives, including *Dasyphyrum villosum*, *Thinopyrum intermedium*, rye, and various *Aegilops* species, have shown effective resistance or immunity against powdery mildew (Chen, Shi, Shang, Leath, & Murphy, 1997; Friebe, Heun, Tuleen, Zeller, & Gill, 1994; Jia et al., 1996; Miranda, Murphy, Marshall, & Leath, 2006; Shubing & Honggang, 2005). Various PM resistance genes, including *Pm12*, *Pm13*, *Pm16*, *Pm20*, *Pm4b* and *Pm21*, have been utilized from *A. speltoides*, *Aegilops longissima*, *Triticum dicoccoides*, *Secale*, *Aegilops Ventricosa*, and *Heterotheca villosa*, respectively (Zhou et al., 2005). *Pm12*, *Pm21*, and *Pm37* impart high resistance against powdery mildew, with *Pm21* derived from the *Triticeae* grass *D. villosum* conferring high immunity to various pathogen races (He et al., 2018).

Till date, identifications of 88 formally designated *Pm* genes or alleles have been done at 66 loci (*Pm1–Pm66*) (Li, Dong, et al., 2020; Zhu et al., 2020). Additionally, more than 30 temporarily designated *Pm* genes are reported and assigned to their corresponding wheat chromosomes (Li, Shi, et al., 2020). There are two types of PM resistance reported in wheat, the one mediated through the use of resistance genes while the other resistance mediated through the mutation negative regulators of PM resistance, like *Mildew Resistance Locus (MLO)* and *ENHANCED DISEASE RESISTANCE 1 (EDR1)* (Zou, Wang, Li, Kong, & Tang, 2018). Many PM resistance genes identified and mapped in wheat and among them *Pm2*, *Pm3*, *Pm8*, and *Pm21* had been cloned (Cao et al., 2011; Hurni et al., 2013; Sánchez-Martín et al., 2016; Yahiaoui et al., 2004). The best known and most significantly utilized gene is *Pm8* which has played an important role for protecting the wheat yield loss from PM infection. This gene was transferred from the “Petkus” rye chromosome into hexaploid wheat in the early 1930s. In addition to powdery mildew resistance, the rye chromosome arm 1RS offers resistance to other diseases such as strip rust. Recently, Tang, Hu, Zhong, and Luo (2018) investigated the potential role of *Pm40* in Chinese wheat breeding programs in the post-*Pm21* era. Since, new pathogen isolates virulent to *Pm21* have been identified in some wheat fields, as such the use of *Pm40* is believed to offer a durable and broad spectrum of resistance to the newly evolved pathogenicity.

19.3.3 Genomic hotspots for drought stress tolerance

Balancing plant yield and growth in a drought-stressed area is the prime objective for wheat breeding programs. Controlled by several genes, drought tolerance is a quantitative trait with low heritability. The abiotic stress triggers expression of various genes, with effect on the metabolism of different major enzymes, hormones, carbohydrates, and transcription factors. The remarkable ones comprise abscisic acid (ABA), tryptophan, raffinose, late embryogenesis abundant (LEA) proteins, superoxide dismutase, and glycine betaine (Hameed, Bibi, Akhter, & Iqbal, 2011; Nio, Cawthray, Wade, & Colmer, 2011; Sivamani et al., 2000). These play major role in the events for avoiding dehydration like adjusting the osmotic balance, antioxidant effect, and functioning as scavengers for reactive oxygen species. Implementation of information acquired on signaling and metabolic processes in which these biomolecules are involved has led to improvement in drought-tolerant crop species through transgenic approach. Sivamani et al. (2000) introduced the ABA-responsive barley gene *HVA1*, a member of group three LEA protein genes into spring wheat through the biolistic method. Results indicated improvement in growth characteristics in these wheat lines with enhanced water use capacity, root weight, and biomass accumulation in response to deficit soil water. Vendruscolo et al. (2007) in their work studied role of proline-inducing gene (*P5CS*) in boosting drought tolerance of wheat transgenic lines, which was predominantly through protective mechanism against the oxidative stress. Accumulation of proline indicated positive approach in maintaining the productivity of plants under such stress situations. In their investigation, used in another study, *mtlD* gene from *Escherichia coli* was used for the biosynthesis of mannitol to improve wheat tolerance against salinity and water stress. Results of the study concluded increased growth of calli-accumulating mannitol and mature leaves because of stress-protective character of mannitol.

Around 1200 QTLs have already been reported for different drought-responsive traits in wheat (Gupta, Balyan, Sharma, & Kumar, 2020; Gupta, Rico-Medina, & Caño-Delgado, 2020; Kumar et al., 2020). Root system architecture (RSA) is an essential target for breeding wheat with drought tolerance (Lopes & Reynolds, 2010). Roots with greater surface area impart the plant with potential to take up more water and nutrients. Process of soil–root interaction is also somewhat explained by the root angle trait (Chen et al., 2017; Chen, Li, He, & Ding, 2018). Nodal and seminal roots having narrower root angles are likely to grow deeper in the soil in comparison to the wider root angle (Manschadi, Hammer, Christopher, & Devoil, 2008; Richard et al., 2015; Wasson et al., 2012). Moreover, denser lateral roots with narrow-angle are regarded well because these roots are more accessed to soil moisture in deeper soil. Root angle is highly heritable and therefore it is a major trait for consideration during wheat breeding (El Hassouni et al., 2018). Sharma et al. (2011) found major QTL linked to *Pm8* in 1RS for enhancing root biomass for better water use efficiency and uptake of nutrients, thus providing drought tolerance to the growing crop.

Genomic regions related to various physiological traits have been studied like chromosome 2D and 2B for flag leaf senescence (Verma et al., 2004); chromosome 6A for seedling vigor (Spielmeyer et al., 2007); 1B, 2B, 3B, 4A, and 5A for canopy temperature (Pinto et al., 2010); and 6A and 4B for coleoptile length (Rebetzke et al., 2001). Merchuk-Ovnat et al. (2016) demonstrated enhanced drought tolerance by introgressing QTLs on 1B and 2B of *Triticum turgidum* into *T. aestivum*. Zandipour, Hervan, Azadi, Khosroshahli, and Etminan (2020) identified

QTL on chromosome 1B for nine important traits under terminal drought stress conditions in wheat. [Maulana, Huang, Anderson, and Ma \(2020\)](#) studied significant QTLs associated with seedling drought tolerance—related traits on 1B, 2A, 2B, 2D, 3A, 3B, 3D, 4B, 5A, 5B, 6B, and 7B. Out of these, 12 stable QTLs responding to drought stress for various traits were identified. [Tura et al. \(2020\)](#) worked on analyzing and mapping of QTLs for yield-related traits under drought conditions on chromosome 4A, 5B, and 7A. Later, [Gautam et al. \(2020\)](#) efficiently introgressed a yield QTL (Qyld.csdh.7AL) into four elite Indian wheat cultivars for developing drought-tolerant genotypes.

19.3.4 Genomic hotspots for heat tolerance in wheat

Heat stress reduces the yield of wheat by 33.6%, attributed to reduced chlorophyll content, respiration rate, photosynthesis, and dehydration lead seedling death ([Cossani & Reynolds, 2012](#)). The significant effect of heat on the plant is reduced photosynthesis because of early leaf senescence and decrease in leaf area expansion which ultimately leads to reduced grain production ([Mathur, Agrawal, & Jajoo, 2014](#)). In tissues involved in photosynthesis, photosystem-II is largely responsible for heat stress where increased fluidity of thylakoid membrane and electron transportation in response to heat stress is observed ([Prasad, Pisipati, Ristic, Bukovnik, & Fritz, 2008](#)). Stress caused due to increased temperature at the grain filling stage leads to reduction in yield and quality of wheat ([Maulana et al., 2020](#)). This increased temperature during the time of grain filling is known as terminal heat stress.

Certain parameters for indicating heat tolerance include canopy temperature, normalized difference vegetation index, and chlorophyll content ([Hazratkulova et al., 2012](#)). Moreover, stay green trait has been effectively used for evaluating heat tolerance. This trait permits the plant in retaining the photosynthetic ability till advanced stages of the plant under heat stress, enhancing the grain filling period and ensuring yield stability ([Kumari, Pudake, Singh, & Joshi, 2013](#)). Ground cover is an indication of genotype efficiency in producing canopy area and biomass which can be estimated through digital imaging and canopy reflectance indices ([Mullan & Reynolds, 2010](#)). Additionally, canopy temperature depression and cell membrane stability are useful traits in identifying donors at an early stage ([Reynolds, 1997](#)).

[Sarieva, Kenzhebaeva, and Lichtenthaler \(2010\)](#) observed that leaf rolling is helpful in stabilizing the organization of PSII and PSI under short-lived heat stress. The extent of leaf rolling helps in determining the sustenance of optimal photosynthetic activity in leaves. As such leaf rolling provides high adaptability facilitating water metabolism in flag leaves efficiently. Using molecular markers with other physiological characteristics helps in identifying elite germplasms, new alleles and making better improvements for heat tolerance. [Sharma et al. \(2014\)](#) found four highly heat-tolerant lines among 24 synthetic wheat lines. In general, *T. monococcum* and *T. dicoccoides* are better options as germplasm to increase heat tolerance in bread wheat. Moreover, it was also found in *A. longissima*, *Aegilops searsii*, and *A. speltoides* ([Choudhary, Yadav, & Saran, 2020](#)). *A. speltoides* is a prime genetic resource for the improvement of this trait in wheat. [Awlachev, Singh, Kaur, Bains, and Chhuneja \(2016\)](#) crossed *A. speltoides* PAU 3809 with *Triticum durum* cv. PDW274 to develop backcross introgression lines. Genotyping using SSR markers and mapping the QTL controlling the heat-tolerant trait was done for these lines. Also, phenotyping was done for acquired thermo-tolerance, chlorophyll content, canopy temperature, stay green, and cell membrane thermo-stability. QTLs for different heat-tolerant traits were identified on 2B, 3A, 3B, 5A, 5B, and 7A. [Yang, Sears, Gill, and Paulsen \(2002\)](#) identified two markers, Xgwm293 and Xgwm11 related to grain filling for F₂ population with findings demonstrating that heat tolerance of common wheat is controlled by multiple genes. Furthermore, QTL for canopy temperature during heat stress has been investigated to coincide with QTL for yield and heat susceptibility index ([Mason, Mondal, Beecher, & Hays, 2011](#)).

Using metaanalysis, important QTLs related to heat tolerance were found on 1B, 2B, 2D, 4A, 4D, 5A, and 7A ([Acuña-Galindo, Mason, Subramanian, & Hays, 2015](#)). The Langdon chromosome substitution lines were initially used for mapping genes involved in heat tolerance in the year 1991 and were found on chromosomes 3A, 3B, 4A, 4B, and 6A ([Ni et al., 2018](#); [Sun & Quick, 1991](#)). Later, [Ruqiang, Qixin, and Shuzhen \(1996\)](#) in their study concluded that chromosomes 3A, 3B, and 3D were related to heat tolerance in cultivar Hope. Also, heat-tolerant sources like NIAW 845, WH 730, RAJ 4037, HD 2808, and NIAW 34 have been used in breeding wheat. Moreover, RAJ 3765, HW 2045, Lok 54, RAJ 4250, WH 1021, HD 3095, and GW 432 were identified for having least heat-sensitivity index under Multilocation Heat Tolerance Trial. As such these genotypes have potential for improving wheat against heat stress ([Mishra et al., 2014](#)).

19.4 Genomic sequences to genomic hotspot

Wheat-genome sequences were collected, indexed, and organized in the International Wheat Genome Sequencing Consortium (IWGSC) data repositories hosted by URGI (Unité de Recherche Génomique Info/research unit in genomics and bioinformatics), INRA (Institut National de la Recherche Agronomique/French National Institute for Agricultural Research) to ensure FAIR principle, that is, data should be Findable, Accessible, Interoperable, and Reusable. The basic pipeline for genome sequence analysis involves primary, secondary, and tertiary analyses. The primary analysis includes analysis of hardware-generated data and machine stats. This covers production of sequence reads and quality scores. The secondary analysis includes quality analysis (QA) filtering on raw reads, alignment/assembly of reads, followed by QA and variant calling on aligned reads. The tertiary analysis is mostly objective driven, includes multisample processing, QA/QC of variant calls, annotation and filtering of variants, data aggregation, association analysis, population structure analysis, genome browser driven exploratory analysis (Malviya, Yadav, & Yadav, 2019). Furthermore, to make sense out of the genome sequences, different analyses are performed such as sequence similarities and homologies, identification of sequence features such as gene structure, distribution, introns–exons, and regulation of gene expression. Additionally, sequence variations such as insertion, deletions, and single-nucleotide polymorphism (SNP) are key for gene–trait association. The availability of wheat reference genome provides comprehensive information about genes and genetic factors on different chromosomes. However, contribution of these factors to adaptive traits (heat and drought tolerance, rusts/yellow-spot/powdery mildew resistance) and agronomic traits (grain quality and yield) has been governed through different approaches. The genomic regions with genetic factors contributing to the adaptive and agronomic traits are considered to be genomic hotspots where the small regions of genome controlling the major adaptive traits. The exploration and inclusion of these genomic hotspots is the major objective of modern next-generation breeding. The SNP array and diversity array technology (DArT) platforms are the potential tools with high density, high-throughput, and low-cost marker system used for marking the genomic hotspots (Gupta, Langridge, & Mir, 2010; Rasheed, Mujeeb-Kazi, Ogbonnaya, He, & Rajaram, 2018). The SNP offers locus specificity, codominance, high-throughput, and comparatively low genotyping errors (Rafalski, 2002). However, limitation of lack of flexibility and ascertainment bias are reported for SNP array-based genotyping (Albrechtsen, Nielsen, & Nielsen, 2010; Thomson, 2014). SNP markers are fixed on the array and used in a defined way; however, if additional SNPs are required, the array need to be redesigned, which can be expensive.

The IWGSC reference wheat-genome sequenced data provide wheat breeders with the ultimate choice and precise outcomes, which will continue to make a huge impact on all aspects of wheat improvement in years to come (IWGSC, 2018). However, the raw sequence information to hotspots in wheat required object-oriented customized tools. These involved mainly sequence alignment and assembly tools (Table 19.1). The alignment-based tools such as BLAST, FASTA, multiple sequence aligners (e.g., ClustalW, Muscle, MAFFT), sequences profile search programs (e.g., PSI-BLAST, HMMER/Pfam), and whole-genome aligners (e.g., BLASTZ, TBA) are dependent on colinearity. They rely on dynamic programming, requiring huge computational memory and time costing money and labor. Further, these approaches are not effective under the genetic shuffling and recombination events. Also, these approaches are based on assumptions about evolution of the sequences, including various parameters such as substitution matrices, gap penalties, and threshold values (Misale, Ferrero, Torquati, & Aldinucci, 2014).

Alignment-free methods are developed to overcome these challenges. First, alignment-free method is based on the frequencies of subsequences of a defined length, also known as word-based methods. Second, it evaluates the informational content between full-length sequences and known as information theory–based methods. Furthermore, methods were developed based on the length of matching words (common, longest common, or the minimal absent words between sequences), chaos game representation, iterated maps, as well as graphical representation of DNA sequences, which capture the essence of the base composition and distribution of the sequences in a quantitative manner. All these methods are well supported mathematically, statistically and calculate pairwise measures of dissimilarity or distance between sequences. Further, these measures can be directly used in standard tree-building tools such as Phylip and MEGA (Zielezinski et al., 2017).

In frequency-based methods, similar sequences share similar words/k-mers (subsequences of length k), and mathematical operations with the words' occurrences give a good relative measure of sequence dissimilarity. As a rule of thumb, smaller k -mers should be used when sequences are not related whereas longer k -mers can be used for very similar sequences. These methods are operated on vectors, thus allow the use of more than 40 functions other than the Euclidean distance such as Pearson correlation coefficient and Manhattan distance (Vinga & Almeida, 2013).

TABLE 19.1 Common objective oriented tools used to predict genomic hotspots.

Common general sequence analysis tools	
1	ALP (Ascending Ladder Program) to calculate the statistical parameters for BLAST.
2	Arioc is a set of tools to align short bisulfite-treated DNA sequences (BS-seq reads) to long reference DNA sequences.
3	BuddySuite is a collection of four related tools: <i>SeqBuddy</i> is a tool to handle FASTA, GenBank, and NEXUS sequence file formats. <i>AlignBuddy</i> : 30 separate tool modules to read, write, analyze, and manipulate PHYLIP, Stockholm, and NEXUS sequence alignment files. <i>PhyloBuddy</i> : Consists of 18 tool modules to manage and manipulate phylogenetic trees in NEXUS, Newick, and NeXML formats. <i>DatabaseBuddy</i> : Contains function to search NCBI, UniProt, and Ensembl databases.
4	CorGen is a web-based tool to measure long-range correlations in DNA sequences characterized by a power-law decay of the autocorrelation function of the GC-content.
5	cpgplot is a tool for plotting and identification of CpG islands in nucleotide sequences.
6	DAMBE7 is a tool for genomic and phylogenetic sequence data analysis.
7	PyBamView is a tool to visualize sequence alignments from BAM files with an optional of FASTA-formated reference genome.
8	SPARSE (Sparsified Prediction and Alignment of RNAs based on their structure Ensembles) is a tool to align RNA sequences based on structural properties of RNA ensembles.
9	supermatcher is a tool to compute approximate alignments between search sequences and the target sequences.
10	WebSat is a web-based tool to predict molecular markers, visualization of microsatellites, and design primers for them.
Repeat analysis tools	
1	CHOPCHOP is a web-based tool to select target sites for CRISPR/Cas9- or TALEN-directed mutagenesis.
2	CRISPRCasFinder is a tool to find CRISPR (clustered regularly interspaced short palindromic repeats) arrays and detect Cas proteins.
3	CRISPRFinder is a web-based tool for the discovery of CRISPRs, the definition of direct repeats (Dr), extraction of spacers, obtaining flanking sequences from the Genbank database.
4	detectIR is a tool to find perfect and imperfect repeats and inverted repeats in DNA sequences.
5	Dfam is a web-based database containing transposable element DNA sequence alignments (interspersed repeats), Hidden Markov models (HMMs), consensus sequences, and genome annotations.
6	etandem is a tool to find tandem repeats in DNA sequences.
7	einverted is a tool for finding inverted repeats, or stem-loops, in nucleotide sequences.
8	HipSTR (Haplotype inference and phasing for Short Tandem Repeats) is a tool to genotype, phase short tandem repeats (STRs), and to analyze and validate de novo STR mutations genome-wide.
9	Kmer-SSR is a tool to detect simple sequence repeats (SSRs) in genomic sequences.
10	lobSTR is a tool to align and genotype short tandem repeat profiles from high-throughput sequencing data.
Whole genome analysis tools	
1	A5 is a tool for automating genome assembly pipeline and it consists of five steps: cleaning reads, assemble error-corrected reads, scaffolding, scaffold validation, and final scaffold assembly.
2	ABYSS is a tool for de novo genome assembly using short-read data. It implements a distributed representation of de Bruijn graphs, which enable parallel computation of the assembly algorithm. ABYSS stands for Assembly By Short Sequencing.
3	ALLPATHS is a tool for genome assembly that is applicable to all types of sequences and not limited to just short reads.
4	AutoSeqMan is a tool for assembling Sanger sequences into contigs for users working with the Seqman program.
5	BioNanoAnalyst is a tool for evaluating potential misassemblies in reference genomes using optical maps.

(Continued)

TABLE 19.1 (Continued)

6	CANU is a tool to assemble long reads from either PacBio or Oxford Nanopore, which have higher error rates than short reads from Illumina.
7	Celera is the first generation of assembler capable of assembling the genomes of multicellular organisms. It was used to assemble the model organism, fruit fly, and subsequently used to assemble the first human genome.
8	Edena is a tool for de novo genome assembly that is based on overlap layout assembly framework and it is applicable to very short reads from the Illumina platform (e.g., 35 bp).
9	ELOPER is a tool to preprocess paired-end short reads for a better performance during assembly. It implements an algorithm that detects overlaps between both ends of the paired-end reads, which then merged those reads with significant overlaps.
10	FALCON is a tool for de novo assembly of long PacBio reads and it is an improved version of its predecessor HGAP. Unlike HGAP, it is a diploid-aware assembler that is better suited to assemble larger genomes.
11	Hapler is a tool for assembling sequences into haplotypes from population-sampled data.
12	HapTree is a tool for haplotype reconstruction from sequencing data (e.g., Illumina) of a single individual genome that may be diploid or has higher ploidy.
13	Kermit is a tool for using linkage maps to guide genome assembly.
14	laSV is a tool for detecting structural variants (SVs) from paired-end sequenced data at single base pair resolution.
15	LightAssembler is a lightweight program for genome assembly based on the use of a pair of cache-oblivious Bloom filters.
16	MaSuRCA is a tool for genome assembly based on a hybrid approach that combines de Bruijn graph and overlap-based assembly strategies. It can be used for sequenced data with variable read lengths and hence it is suitable for assembling 454, Sanger and Illumina data.
17	MindTheGap is a tool specifically designed to assemble insertion variants from resequencing data.
18	miniasm is a tool for de novo assembly of long reads from either the PacBio or Oxford Nanopore platforms.
19	misFinder is a tool for checking assembly errors by using a reference genome and alignments of paired-end reads.
20	misSEQuel is a tool for detecting and correcting errors in draft assemblies.
21	npScarf is a tool for assembly scaffolding and gap filling suitable for smaller genomes already assembled with short reads.
22	Orione is a Galaxy-based framework that grouped together workflows and tools to perform de novo genome assembly, annotation, RNA-Seq, and metagenomics analysis.
23	PASHA is a tool for assembling genomes based on short reads using the de Bruijn graphs with the main improvement being its code for distributed computing.
24	Ray Meta is a tool for de novo assembly of metagenomes using distributed computing to enable parallel assemblies of multiple genomes.
25	SHORTY is a tool for de novo genome assembly of short reads, in particular reads generated from the SOLiD sequencing platform.
26	SMRT is a tool for calling single-nucleotide polymorphisms (SNPs) and assembling haplotypes based on long PacBio reads. The name for this tool is based on Single Molecule Real Time (SMRT) sequencing and the paper describing this tool used PacBio reads.
27	SOAPdenovo is a tool for de novo genome assembly using entirely Illumina short reads. The algorithm implements error correction, de Bruijn graph construction, tip removal, repeat resolution, bubbles merging, contig linkage graph, and scaffolding.
28	SQUAT is a tool for both preassembly and postassembly evaluation. The preassembly evaluation is based on read quality whereas the postassembly steps take into account how well reads are mapped onto a reference genome.
29	Velvet is a tool for de novo assembly based on de Bruijn graphs and it is suitable for short-read data with high coverage.
30	WhatsHap is a tool for phasing long reads despite their higher sequencing error rates. It implements a fixed-parameter tractable (FPT) approach to a weighted version of minimum error correction (wMEC) formulation. This tool is useful to users who want to perform haplotype assembly.
Genome-wide association study (GWAS) tools	
1	AlphaDrop beta is a tool to simulate genomic selection and GWAS data. It can simulate sequence data, SNP data, pedigrees, QTL effects, and breeding values.

(Continued)

TABLE 19.1 (Continued)

2	BLINK (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway) is a tool for GWASs to identify genes controlling human diseases and agricultural traits. The algorithm uses Bayes and linkage disequilibrium information.
3	BioBin is a tool to investigate the rare variant burden in genetic trait studies and the natural distribution of rare variants in ancestral populations aimed for analyzes and hypothesis testing. The BioBin algorithm generates bins based on various features, such as regulatory and evolutionary conserved regions, genes, and pathways. BioBin uses information stored in the LOKI database, the Library of Knowledge Integration.
4	EMMAX (Efficient Mixed-Model Association eXpedited) is a tool for testing association mapping considering the sample structure in GWASs. The EMMAX algorithm uses a variance component approach that can analyze GWAS datasets within hours.
5	EPIQ is a tool to detect epistasis in quantitative GWAS. The EPIQ algorithm uses metric embedding and random projections to eliminate the need to exhaustively test all SNP pairs.
6	G2P (A Genome-Wide-Association-Study Simulation Tool for Genotype Simulation, Phenotype Simulation, and Power Evaluation) is a tool to simulate genotypes for the GWASs. The G2P can simulate genotype data, phenotype data and evaluate the statistical power.
7	GAPIT (Genome Association and Prediction Integrated Tool) is a tool for GWAS and genome prediction or selection. The GAPIT algorithm uses The Mixed Linear Model (MLM).
8	GEMMA (genome-wide efficient mixed-model association) is a tool for testing association in GWAS data. The GEMMA algorithm computes exact Wald statistics and <i>P</i> -values.
9	GIGSEA (Genotype Imputed Gene Set Enrichment Analysis) is a tool to analyze imputed genotypes. The GIGSEA algorithm uses a combination of GWAS summary statistics and eQTL to deduce differential gene expression and to examine enrichment for gene sets.
10	GenoWAP is a tool to prioritize signals, integrate functional annotation, and GWAS test statistics in GWAS results.
11	GPA (Genetic analysis incorporating Pleiotropy and Annotation) is a tool for the prioritization of GWASs results using pleiotropy information and annotation data. The GPA algorithm has functions for fitting models and hypothesis testing the associated SNPs.
12	GWAS catalog is a catalog of publicly available, manually curated, and published GWAS data, containing over 100k SNPs and trait associations.
13	GWAS Pipeline is a pipeline tool for genome-wide association analysis (GWAS). The GWAS pipeline can filter, create a kinship matrix, covariate files, run EMMAX, computes Manhattan and QQ plots. The GWAS has functions for computing a summary of the most significant SNPs with calculated allele effects.
14	GWASTools is an R tool for quality control, analysis, and annotation of GWAS data. The package stores data in NetCDF format to allow datasets that exceed the R memory limits.
15	HaploView is a tool to analyze and visualize LD haplotype maps. The HaploView includes functions for LD and haplotype block analysis, population frequency estimation, single SNP and haplotype association tests, permutation testing, Paul de Bakker's Tagger tag SNP selection algorithm, download of phased genotype data from HapMap, visualization, and plotting.
16	TASSEL , A tool to evaluate trait associations, linkage disequilibrium, and evolutionary patterns.
Single-nucleotide polymorphism (SNP) analysis tools	
1	ASSIST is an automatic SNP scoring tool for in- and outbreeding species. Customized pipeline for calling and filtering of SNPs from Illumina Infinium arrays. ASSIST builds on GenomeStudio-derived data and identifies markers that follow a biallelic genetic model and show reliable genotype calls, and re-edits SNP calls.
2	ALICE an integrated analysis of allele frequency, allelic imbalance, loss of heterozygosity, long contiguous stretch of homozygosity, and copy number variation or alteration based on SNP probe hybridization intensities and genotypes.
3	ATLAS-SNP2 is a SNP discovery method to assess variant allele probability.
4	Chopsticks is an R tool containing classes and methods for large-scale single-nucleotide (SNP) association studies.
5	ChroMos , SNP classification, prioritization and prediction of their functional effect. The tool uses a large database of SNPs and chromatin states, and allows a user to upload genetic information.
6	CircosVCF , a visualization tool of genome-wide variant data described in VCF files using circos plots. Gives a broad overview of the genomic relationship between genomes, and can focus on specific SNP regions.
7	CsSNP , detection of comparative SNP segments and display detailed information of them.

(Continued)

TABLE 19.1 (Continued)

8	DnaSP , a software package for the analysis of DNA polymorphisms using data from a multiple sequence aligned data. Features: Estimate various measures of DNA sequence variation within and between populations. Estimate linkage disequilibrium, recombination, gene flow and gene conversion parameters.
9	Ebwt2snp , a tool to detect SNP using extended Burrows–Wheeler Transform (eBWT), and makes a reference-free evaluation of its accuracy by calculating the coverage of each SNP.
10	SNP Effect Predictor , a web and command-line tool to analyze and predict functional consequences of your variants, SNPs, indels, CNVs on genes, regulatory regions, transcripts, and protein sequences.
11	Fast-GBS , a pipeline to extract a high-quality SNP catalog. INPUT; FASTQ files obtained from sequencing genotyping-by-sequencing (GBS) libraries.
12	FastSNP , a web-based tool for the identification of tumor-associated SNP.
13	Flapjack , a tool for analysis and visualization of large volumes of SNP. The real-time graphical rendering allows comparison between lines, markers, and chromosomes.
14	FunctSNP , a R package to link SNP to function knowledge. dbAutoMaker generates a local database.
15	Heap , SNP calling in low coverage NGS data, aligned to the reference genome sequences. Heap determines genotypes and calls SNPs.
16	InSNP , a tool for SNP and indel detection.
17	ISMU , a pipeline integrating several open source next-generation sequencing (NGS) tools along with a graphical user interface called Integrated SNP Mining and Utilization (ISMU) for SNP discovery and their utilization by developing genotyping assays.
18	KASPSpoon , a tool for high-throughput SNP genotyping.
19	kSNP , finds SNPs in whole-genome data. kSNP v2 has numerous improvements over kSNP v1 including SNP gene annotation.
20	Mapsnp , the tool uses the biomaRt and the rtracklayer packages to annotate queries to Ensembl and UCSC and translates this to, for example, gene/transcript structures in viewports of the grid graphics package. mapsnp package inherits from the Gviz package.
21	MapNext , an automated SNP detection from population sequences for spliced and unspliced alignments of short reads.
22	mrSNP , Software to detect SNP effects on microRNA binding.
23	MSQT , a common-line tool to extract SNP data from multiple sequence alignments, stores it in a database, and provides a web interface to query the database.
24	NASP , a tool to identify SNP in whole genome sequencing data
25	NovelSNPer , a web tool to classify sequence variants based on the gene structure information in Ensembl.
26	QQ-SNV , a tool to detect SNP in heterogeneous virus population from Illumina sequencing data. The QQ-SNV algorithm uses a logistic regression classifier model based on quantiles of quality scores.
27	Seq-SNPing , a SNP discovery, ID identification, editing, and visualizing of sequence alignments.
28	SCcaller , a tool for detection of SNP and short insertions and deletions (indels) in data from single-cell sequencing.
29	SMuRF , R package for selection of regulatory elements, single-nucleotide variants (SNVs), SNPs, and is-regulatory elements (CREs) using random forests.
30	SNIP-Seq , a tool to detect SNP in Illumina sequence data from population samples. The SNIP-Seq algorithm uses quality values of the sequenced bases and iterative estimates of genotypes and error rates based on multiple individuals.
31	SNiPlay3 , a web tool for pipelines to detect, manage, and analyze SNPs and indels.
32	SNiPloid , a tool to discover and validate predicted SNPs, optimized for allopolyploid species.
33	SNPchip , R package to plot SNP data.
34	SNPdetector , an automated identification of SNPs and mutations in fluorescence-based resequencing reads (Sanger sequencing reads).
35	SNPHarvester , a tool to search for significant SNP groups in large-scale association studies. It can select a set of significant SNP groups from hundreds of thousands of SNPs efficiently.
36	SNPmeta , SNP annotation and SNP metadata collection of nonmodel species or species that lack a reference genome.

(Continued)

TABLE 19.1 (Continued)

37	SNPs3D , a website with tools for assigning molecular functional effects of non-synonymous SNPs based on structures and sequences.
38	SNPServer , a web tool for discovery of SNPs within DNA sequence data. The program uses BLAST, to identify related sequences, and CAP3, to cluster and align these sequences. The alignments are parsed to autoSNP, a program that detects SNPs and insertion/deletion polymorphisms.
39	SNPTools pipeline comprises tools for SNP analysis in next-generation sequencing data. It has an imputation engine refining raw genotype likelihoods to output high-quality genotypes or haplotypes, designed for genotyping studies of large populations.
40	SparSNP , a fast and memory-efficient analysis of all SNPs for phenotype prediction. It also does cross-validation.
RNA-seq analysis tools	
1	ABYSS is a tool for de novo genome assembly using short-read data. It implements a distributed representation of de Bruijn graphs, which enable parallel computation of the assembly algorithm. ABYSS stands for Assembly By Short Sequencing.
2	Oases is a tool for assembling de novo transcriptomes using short RNA-seq reads. The Oases algorithm uses dynamic error removal in the prediction of full-length transcripts, and it can handle a wide range of expression values and the absence of alternative isoforms. Requires Velvet 1.2.08 or higher.
3	Trinity is a tool for de novo transcriptome assembly of RNA-seq data and consists of three modules: Inchworm, Chrysalis, and Butterfly. The algorithm uses de Bruijn graphs, dynamic programming method, it can detect isoforms, handle paired-end reads, multiple insert sizes, and strandedness.
4	SOAPdenovo-Trans is de novo RNA-seq full-length transcriptome assembler. The SOAPdenovo-Trans algorithm adapts the SOAPdenovo framework, uses the Trinity error removal technique, the graph traversal model from Oases, and uses a transitive reduction to simplify scaffolding graphs. It can handle paired-end reads and multiple insert sizes.
5	GSNAP (Genomic Short-read Nucleotide Alignment Program) is a tool to align single- and paired-end reads to a reference genome. The GSNAP algorithm is based on the seed-and-extend method and works on reads down to 14 nucleotides of length, and computes SNP-tolerant alignments of various combinations of major and minor alleles.
6	STAR , a tool to align RNA-seq data. The STAR algorithm uses suffix arrays, seed clustering, and stitching. It can detect noncanonical splice sites, chimeric sequences, and can also map full-length RNA sequences.
7	TopHat is a tool for splice-aware mapping of RNA-seq reads. The TopHat uses the Bowtie short-read aligner tool (BWT-based algorithm) for the mapping whereafter it identifies intron-exon (splice) junctions. TopHat can use paired-end sequencing reads and parallel computation.
8	MapSplice is a tool to align RNA-seq read to a reference sequence. The MapSplice algorithm uses the Burrows—Wheeler Transform (BWT) technique and can discover both canonical and noncanonical splice sites.
9	Rbowtie2 is an R tool that wraps the Bowtie 2 tool and includes adapter removal, read merging and identification.
10	Rbowtie package provides an R wrapper around the popular bowtie short-read aligner and around SpliceMap, a de novo splice junction discovery and alignment tool. The package is used by the QuasR bioconductor package. We recommend to use QuasR instead of using this package directly.
11	DeepBound is a tool to identify splicing junctions and boundaries of expressed transcript read alignments in RNA-seq data. The DeepBound algorithm uses deep convolutional neural fields.
12	SpliceJumper is a tool to identify splice junctions in RNA-seq data. The SpliceJumper algorithm uses a classification-based approach.
13	MapPER is a tool to align paired-end reads in RNA-seq datasets. The MapPER algorithm uses an expectation—maximization method to assign likelihood values.
14	NanoPARE is a set of tools for the analysis of 5' RNA data from nanoPARE sequencing libraries.
15	GRIT (Generalized RNA Integration Tool) is a tool to assemble transcripts using RNA-seq data. The GRIT pipeline combines RNA-seq and gene-boundary data, CAGE, RAMPAGE, and poly(A)-seq data.
16	RNA-SeQ , a tool for quality control of RNA-seq data. The RNA-SeQC package has functions for computing various quality metrics, such as alignment quality, duplication rates, GC bias, rRNA content, coverage continuity, covered alignment regions, transcript count, and 3'/5' bias. It produces Read counts, coverage, correlation quality control metrics, and is also suitable for use with scRNA-seq datasets.

(Continued)

TABLE 19.1 (Continued)

17	QualiMap and a later version, Qualimap 2, is a tool for quality control of sequence alignments and genomic features. The QualiMap can use whole-genome and exome sequencing, RNA-seq, and ChIP-seq data. It also has functions for comparison of multiple samples and clustering of epigenomic profiles.
18	Subread is a software tool package for the alignment of both DNA-seq and RNA-seq read data, quantification, and mutation detection. The Subread package consists of five separate tools: (1) Subread, a read aligner for both RNA-seq and DNA-seq data, (2) Subjunc, read aligner for RNA-seq data, detection of exon-exon junctions and gene fusion events, (3) featureCounts, read counting, (4) Sublong, for aligning long reads using the seed-and-vote technique, and (5) exactSNP, a SNP discovery.
19	featureCounts is a tool to quantify RNA-seq and gDNA-seq data as counts. It is also suitable for single-cell RNA-seq (scRNA-seq) data. It supports multithreading. The featureCounts is part of the Subread package
20	easyRNASEq is a tool to quantify RNA-seq expression data.
21	HTSeq is a tool for the analysis of high-throughput sequencing data. It processes reads aligned with HISTAT or STAR and assign expression value counts. The HTSeq is also suitable for the quantification of single-cell RNA-seq data (scRNA-seq).
22	PennDiff is a tool to quantify RNA-seq data. The PennDiff algorithm uses both transcript-based and union-exon methods.
23	Salmon , a tool to quantify transcript expression in RNA-seq data. The Salmon algorithm can correct for GC bias, and it uses “selective-alignment” and massively parallel stochastic collapsed variational inference to achieve high accuracy and speed. It reports transcripts per million mapped reads (TPM).
24	Kallisto , a tool to quantify RNA-seq data. The kallisto algorithm uses a pseudo alignment approach to speed up the alignment procedure. The “pseudo alignment” approach can quantify reads without making actual alignments. Kallisto can handle paired-end and single-end reads. It reports transcripts per million mapped reads (TPM).
25	RSEM (RNA-Seq by Expectation–Maximization) is a tool for the quantification of RNA-seq data. The RSEM algorithm uses the expectation–maximization technique, it can operate with and without a reference, and reports transcripts per million mapped reads (TPM). RSEM scales linearly with the amount of alignment quantity and uses The Bowtie tool for the read alignments.
26	Cufflinks consist of a suite of tools for differential gene expression analysis of RNA-seq data. It assembles aligned reads in a set of transcripts and estimates the relative abundances. The Cufflinks suite consists of the following tools: cufflinks, cuffcompare, cuffmerge, cuffquant, cuffdiff, and cuffnorm.
27	eXpress is a tool to quantify RNA-seq data, but it is also applicable to ChIP-seq, metagenomics, and large-scale sequencing data in general. The eXpress streaming algorithm computes sequenced DNA or RNA in real-time
28	Solas is a tool to predict and quantify expressed isoforms within observed coding regions in RNA-seq data
29	Rcount is a tool to quantify the number of reads mapped to a specific gene (feature counts) in RNA-seq datasets. The Rcount algorithm specifically addresses the issue arising from reads mapping to multiple locations.
30	MMSEQ is a tool to estimate isoform in RNA-seq data. The MMSEQ algorithm uses a new statistical method that deconvolves the mapping of reads to haplotype-specific isoforms and works with paired-end reads.
31	DESeq is a tool for hypothesis testing and differential gene expression analysis of RNA-seq data. The DESeq algorithm applies the negative binomial distribution and a Likelihood Ratio Test (LRT), it normalizes data by trimmed mean of M-values and circumvents a small sample size by incorporating information from all genes in a set of samples.
32	edgeR is a tool for differential expression (DE) analysis of RNA-seq, ChIP-seq, CAGE, and SAGE data with biological replicates. The edgeR algorithm uses information from all the genes, computes the dispersion using a weighted likelihood and F-test techniques. For the normalization, it can use the trimmed mean of M-values, upper quartile (UQ) procedure, Relative Log Expression (RLE), and DESeq. It can compare two groups, paired and unpaired, or use a Generalized Linear Model (GLM). The upper quartile (UQ) procedure is also applicable to single-cell RNA-seq (scRNA-seq).
33	SARTools is an R tool package for differential expression analysis of RNA-seq data. SARTools uses DESeq2 and edgeR. The input consists of raw count data, experimental description files. It will then normalize, estimate dispersion, and analyze differential gene expression.
34	Cuffdiff 2 is a tool to estimate differential expression at gene and transcript levels. It uses a negative binomial model, normalizes the relative log expression method implemented in DESeq, Inter-sample normalization method Q, and reports Fragments per kilobase million Reads per million mapped reads (FPKM). Cuffdiff 2 is a part of the Cufflinks suite of tools.
35	GOEAST (Gene Ontology Enrichment Analysis) is a Gene Ontology (GO) enrichment analysis tool. It can identify overrepresented GO terms and uses several different data sources and species.

The information theory-based methods recognize and compute the amount of information shared between two analyzed biological sequences. The organization of nucleotide and amino acid sequences is digitally analyzed and translated via information theory tools, such as complexity and entropy. In these methods, a low-complexity sequence (e.g., AAAAAAAAAA) will have smaller entropy than a more complex sequence (e.g., ACCTGATGT). The methods applied in calculation of block entropies and coverage in global sequence analysis and predict transcription factor binding sites, sequences as time series, and entropic profiles in local genome analyses. Additionally, higher level correlations in gene mapping and protein–protein interaction networks are also achieved using these methods (Vinga, 2014).

The evolution of sequencing technologies leads to design of new alignment tools based on different datasets such as short, long, and high-quality reads or sequencing platform-specific datasets (Schadt-Metzker, 2010). Since genomic and transcriptomic sequencing generate large volume of data, alignment required intensive parallel processing and multicore platforms. Most of the tools are based on Smith–Waterman algorithm which is known to be computationally expensive and uses multithreading. Till now, more than 200 assembler tools are designed (Table 19.1); however, none of the assembly or alignment approach can alone reconstruct the genome completely from sequence read data (John & George, 2018). The first common wheat (variety Chinese Spring (CS)) genome sequence was obtained from genome sequence assembling from different platform and related progenitors. In this, Roche 454 pyrosequencing (GS FLX Titanium and GS FLX1 platforms) provides reads up to 500 bp covering 5X depth in 16 Gb genome size. Additionally, the gaps and low coverage regions were tackled by additional sequences such as SOLiD CS short reads, Illumina reads from *T. monococcum*, 454 sequences of *A. tauschii* (D genome donor), and cDNA sequences from *A. speltoides* (B genome donor) using tools such as Newbler (Roche commercial software) and MetaSim. The resulted genome was highly fragmented and predicted to contain 95,000 genes (Guan et al., 2020).

IWGSC wheat-genome sequencing project follows chromosome-based BAC by BAC sequencing on Roche 454 and Illumina Hiseq 2000 platform and used different assembly tools. SOAPdenovo version 1.05 and de novo ABySS were used to assemble the filtered short reads leading to contigs representing 61% of hexaploid wheat-genome containing 133,090 high-confidence genes and 890,576 low-confidence genes with ORF-like structures and referred to as the IWGSC chromosome survey sequence (CSS) assembly (IWGSC, 2014). However, the development of the advanced long-read Pacific Biosciences (PacBio) sequencing technology has played key role in achieving more refined wheat genome using hybrid assembly techniques (Larsen, Heilman, & Yoder, 2014). MaSuRCA assembly pipeline accommodates both PacBio long error-prone and Illumina accurate short reads of *A. tauschii* leading to Aet_Mr.1.0 assembly. Also, the first near-complete 100x coverage hexaploid wheat (CS42) assembly independent of molecular-genetic map was generated using MaSuRCA pipeline. This involved 7.06 billion Illumina 150 bp paired-end reads and 55.5 million PacBio reads leading to 95.7 million superreads, which further used to generate 57 million mega-reads using the same pipeline. Celera Assembler (v8.3) was used to assemble synthetic mate pairs with mega-reads resulting in Triticum 1.0 genome version containing 829,839 contigs and 17.05 Gb size. In another approach, FALCON assembler was used to directly assemble the long reads resulting in FALCON Trit1.0 of 12.94 Gb. Further, these two genome versions were merged using MuMmer, generating final assembly of 15.3 Gb covering nearly complete wheat genome (Guan et al., 2020).

More direct methods were deployed for generating CS genome assembly using optimized data types and specially designed algorithms. This involved scaffolding using assembly program w2rap-contigger on libraries of long mate-pair sequence reads and Tight, Amplification-free, Large insert pair-end Libraries sequences generating almost 3 million contigs (> 500 bp). Further, SOAPdenovo was used to reduce the number of contiguous sequences to 1.3 million and CSS-survey reads were used to anchor scaffolds to chromosomes leading to wheat genome assembly referred to as TGACv1, 13.43-Gb long, representing 78% of the wheat genome. The assembly of final reference wheat genome was significantly supported by genome sequencing of wild emmer wheat, durum wheat, *A. tauschii*, *T. urartu* involving DeNovoMAGIC2 (NRGene), SOAPdenovo2, MaSuRCA, SSPACE tools. The IWGSC RefSeq v1.0 was obtained using DeNovoMAGIC2 assembled whole-genome frame integrating physical maps, GBS data, radiation hybrid maps, BioNano optical maps, and Hi-C data (Avni et al., 2017; Clavijo et al., 2017; Luo et al., 2012).

The information in IWGSC RefSeq v1.0 can be harnessed in different studies and breeding program to locate hotspots for adaptive traits [heat and drought tolerance, rusts/yellow-spot/powdery mildew resistance, and agronomic traits (grain quality and yield)] (Acuña-Galindo et al., 2015; Pandey, Joshi, Bhardwaj, Agarwal, & Katiyar-Agarwal, 2014; Singh et al., 2017; Tang, Xu, Zhao, Wang, & Kang, 2018). Different approaches are used to find genes and their localization on chromosomes for traits that we have already mentioned such as biofortification, biotic stress resistance, and drought stress tolerance. Most recent structural and functional genomics approaches take advantage of bioinformatics

resources and recent long-read sequencing technologies. The information generated has been deployed using genomic repositories and platforms for functional annotation of the wheat genome (Table 19.2). These involved GWA studies, genomic selection, bulk segregant analysis, QTL mapping, RNA-sequencing, exome-sequencing, CRISPR–CAS-mediated gene editing (Bevan et al., 2017; Bhusal, Sarial, Sharma, & Sareen, 2017; Cavanagh et al., 2013). How these approaches provide genes or hotspot region on chromosome summarized in Fig. 19.1. The integration of information from DNA, RNA, protein, metabolite, and phenotype led to identification of genes. The sequencing technologies allow discovery and genotyping of SNP markers which accelerate the exploration of germplasm allelic diversity based on allele mining approach. The genome at DNA and RNA levels provides information about gene function, gene sequences, and genetic factors that underlie complex traits. The genome-wide approaches such as RNA-seq, exome sequencing generate huge volume of “omics” data that provide information about interaction of genes and proteins. Furthermore, genome-editing tools also assist in gene identification and pathways modification. Moreover, the heritability of generated omics data combined with phenotypic variation through genetic marker associations resulting in the genomic regions that control the expression of single or multiple genes (eQTLs), metabolite (mQTLs), and proteins (pQTLs). Molecular breeding and transgenic approaches are used to translate the omics information for crop improvement. In molecular breeding, GWAS is often used to identify the significant relationships between the trait and underlying genetic loci. The identified QTLs are the valuable hotspots of crop improvement and can be introgressed into the elite genotypes to get the desirable output.

TABLE 19.2 A list of available genomic repositories and platforms for functional annotation of wheat genome.

S. no	Resource	URL
Genomic sequence resources		
1.	EnsemblPlants	http://plants.ensembl.org/index.html
2.	Gramene	http://www.gramene.org/
3.	GrainGenes	https://wheat.pw.usda.gov/GG3/
4.	CerealsDB	http://www.cerealsdb.uk.net/cerealgenomics/Index_Home.html
5.	URGI	https://urgi.versailles.inra.fr/
Gene expression resources		
1.	PLEXdb	http://www.plexdb.org/plex.php?database = Wheat
2.	WheatExp	http://wheat.pw.usda.gov/WheatExp/
3.	expVIP	http://www.wheat-expression.com
Variation resource		
1.	Wheat autoSNPsdB	http://autosnpdb.appliedbioinformatics.com.au/index.jsp?species = wheat
2.	T3 Wheat	http://triticeaetoolbox.org/wheat/
3.	KASP markers	http://polymarker.tgac.ac.uk/Markdown?md = DesignedPrimers ; http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/select_using_ideogram.php
Transcription factors (TFs) resources		
1.	wDBTF	http://www.appli.nantes.inra.fr:8180/wDBTF/
2.	WheatTFDB	http://xms.sicau.edu.cn/wheatTFDB/
Genome consortium		
1.	IWGSC	https://www.wheatgenome.org/
2.	Sequencing the <i>Aegilops tauschii</i> Genome	http://aegilops.wheat.ucdavis.edu/ATGSP/data.php
3.	OWWC	http://www.openwildwheat.org/

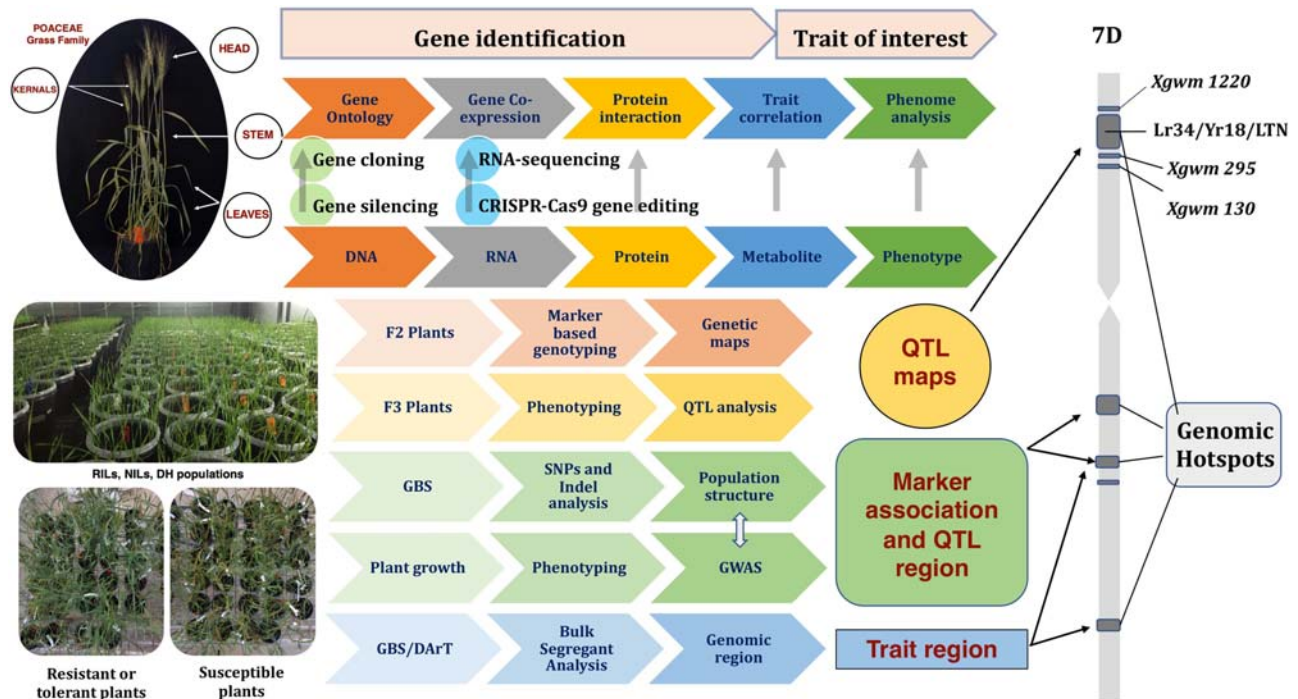


FIGURE 19.1 Representation of genomic sequences to genomic hotspots via different methods and tools such as QTL mapping, genome-wide association studies, genomic selection, bulk segregant analysis, RNA-sequencing, exome-sequencing, CRISPR–CAS9-mediated gene editing where sequenced information harnessed as trait-linked SNPs or insertion/deletion markers. The role of gene in stress tolerance identified using available wheat gene sequences through advanced molecular techniques of gene cloning and gene silencing.

19.5 Conclusion

The application of next generation sequencing (NGS) technology has dramatically expanded our knowledge of genomics especially for wheat where structural and comparative functional genomics would not possible without the advanced high-throughput sequencing technologies and customized bioinformatics pipelines/tools. The long-read PacBio along with Roche 454 and Illumina sequencing has a major impact on accurate sequence assembly of wheat genome. The sequence reads at multiple levels assembled using DeNovoMAGIC2, SOAPdenovo2, MaSuRCA assembly tools resulting in the superior quality wheat reference genome IWGSC RefSeq v1.0. With the availability of genome sequence information discovery and genotyping of the genome-wide markers become feasible. The availability of high-quality genotyping platforms (GBS and DArT-seq) and rich genetic resources opened the new avenues in the molecular breeding likes QTL metaanalysis, conditional QTL mapping, genomic selection, and GBS which has boosted the speed of wheat breeding. The innovation in molecular techniques revealed new approach of CRISPR/cas9, having potential to infer the gene function at precise location and interestingly has the potential to overcome the concern over the use of genetically modified (GM) crops. The recent precise breeding technologies assured the rapid transfer of these hotspots to desired elite wheat line. Finally, the identification and indexing of genomic hotspots for adaptive and agronomic traits will uplift the wheat breeding program worldwide and prepare the wheat crops to take on global climate change.

References

- Acuña-Galindo, M. A., Mason, R. E., Subramanian, N. K., & Hays, D. B. (2015). Meta-analysis of wheat QTL regions associated with adaptation to drought and heat stress. *Crop Science*, 55, 477–492.
- Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, 27, 2534–2547.
- Ali, M. W., & Borrill, P. (2020). Applying genomic resources to accelerate wheat biofortification. *Heredity*, 125, 386–395.
- Alomari, D. Z., Eggert, K., Von Wirén, N., Alqudah, A. M., Polley, A., Plieske, J., ... Röder, M. S. (2018). Identifying candidate genes for enhancing grain Zn concentration in wheat. *Frontiers in Plant Science*, 9, 1313.
- Alomari, D. Z., Eggert, K., Von Wirén, N., Pillen, K., & Röder, M. S. (2017). Genome-wide association study of calcium accumulation in grains of European wheat cultivars. *Frontiers in Plant Science*, 8, 1797.

- Arumuganathan, K., & Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter*, 25, 208–218.
- Asnaghi, C., Roques, D., Ruffel, S., Kaye, C., Hoarau, J. Y., Telismart, H., . . . D'Hont, A. (2004). Targeted mapping of a sugarcane rust resistance gene (Bru1) using bulked segregant analysis and AFLP markers. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 108, 759–764.
- Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science (New York, N.Y.)*, 357, 93–97.
- Awlachew, Z. T., Singh, R., Kaur, S., Bains, N. S., & Chhuneja, P. (2016). Transfer and mapping of the heat tolerance component traits of *Aegilops speltoides* in tetraploid wheat *Triticum durum*. *Molecular Breeding*, 36, 1–15.
- Bailey, R. L., West, K. P., Jr, & Black, R. E. (2015). The epidemiology of global micronutrient deficiencies. *Annals of Nutrition & Metabolism*, 66, 22–33.
- Balance, G., Lamari, L., Kowatsch, R., & Bernier, C. (1998). *The Ptr necrosis toxin and necrosis toxin gene from Pyrenophora tritici-repentis. Molecular genetics of host-specific toxins in plant disease* (pp. 177–185). Springer.
- Balyan, H. S., Gupta, P. K., Kumar, S., Dhariwal, R., Jaiswal, V., Tyagi, S., . . . Kumari, S. (2013). Genetic improvement of grain protein content and other health-related constituents of wheat grain. *Plant Breeding*, 132, 446–457.
- Bancroft, I. (2000). Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana*. *Yeast (Chichester, England)*, 17, 1–5.
- Bennetzen, J. L. (2000). Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *The Plant Cell*, 12, 1021–1029.
- Bertin, I., Zhu, J. H., & Gale, M. D. (2005). SSCP-SNP in pearl millet – A new marker system for comparative genetics. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 110, 1467–1472.
- Bevan, M. W., Uauy, C., Wulff, B. B., Zhou, J., Krasileva, K., & Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature*, 543, 346–354.
- Bhatta, M., Morgounov, A., Belamkar, V., Yorgancilar, A., & Baenziger, P. S. (2018). Genome-wide association study reveals favorable alleles associated with common bunt resistance in synthetic hexaploid wheat. *Euphytica*, 214, 200.
- Bhusal, N., Sarial, A. K., Sharma, P., & Sareen, S. (2017). Mapping QTLs for grain yield components in wheat under heat stress. *PLoS One*, 12, e0189594.
- Bortiri, E., Jackson, D., & Hake, S. (2006). Advances in maize genomics: The emergence of positional cloning. *Current Opinion in Plant Biology*, 9, 164–171.
- Brazma, A., & Vilo, J. (2001). Gene expression data analysis. *Microbes and Infection/Institut Pasteur*, 3, 823–829.
- Brueggeman, R., Rostoks, N., Kudrna, D., Kilian, A., Han, F., Chen, J., . . . Kleinbartsch, A. (2002). The barley stem rust-resistance gene Rpg1 is a novel disease – Resistance gene with homology to receptor kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 9328–9333.
- Brunner, S., Keller, B., & Feuillet, C. (2003). A large rearrangement involving genes and low copy DNA interrupts the microcolinearity between rice and barley at the Rph7 locus. *Genetics*, 164, 673–683.
- Cao, A., Xing, L., Wang, X., Yang, X., Wang, W., Sun, Y., . . . Liu, D. (2011). Serine/threonine kinase gene Stpk-V, a key member of powdery mildew resistance gene Pm21, confers powdery mildew resistance in wheat. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 7727–7732.
- Cavanagh, C. R., Chao, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 8057–8062.
- Ceoloni, C., Kuzmanović, L., Ruggeri, R., Rossini, F., Forte, P., Cuccurullo, A., & Bitti, A. (2017). Harnessing genetic diversity of wild gene pools to enhance wheat crop production and sustainability: Challenges and opportunities. *Diversity*, 9, 1–22.
- Chen, H., Wang, S., Xing, Y., Xu, C., Hayes, P. M., & Zhang, Q. (2005). Comparative analyses of genomic locations and race specificities of loci for quantitative resistance to *Pyricularia grisea* in rice and barley. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 2544–2549.
- Chen, X. (2005). Epidemiology and control of stripe rust [*Puccinia striiformis* f. sp. *tritici*] on wheat. *Canadian Journal of Plant Pathology*, 27, 314–337.
- Chen, X., Ding, Q., Błaszkiwicz, Z., Sun, J., Sun, Q., He, R., & Li, Y. (2017). Phenotyping for the dynamics of field wheat root system architecture. *Scientific Reports*, 7, 37649.
- Chen, X., Li, Y., He, R., & Ding, Q. (2018). Phenotyping field-state wheat root system architecture for root foraging traits in response to environment × management interactions. *Scientific Reports*, 8, 1–9.
- Chen, X., Shi, A., Shang, L., Leath, S., & Murphy, J. (1997). The resistance reaction of *H. villosa* to powdery mildew isolates and its expression in wheat background. *Acta Phytopathologica Sin*, 27, 17–22.
- Cheng, P., & Chen, X. (2010). Molecular mapping of a gene for stripe rust resistance in spring wheat cultivar IDO377s. *Theoretical and Applied Genetics*, 121, 195–204.
- Choudhary, M., Yadav, M., & Saran, R. (2020). Advanced screening and breeding approaches for heat tolerance in wheat. *Journal of Pharmacognosy and Phytochemistry*, 9, 1047–1052.
- Ciuffetti, L. M., Tuori, R. P., & Gaventa, J. M. (1997). A single gene encodes a selective toxin causal to the development of tan spot of wheat. *The Plant Cell*, 9, 135–144.
- Clavijo, B., Garcia-Accinelli, G., Wright, J., Heavens, D., Barr, K., Yanes, L., et al. (2017). W2RAP: A pipeline for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv*, 2017, 110999.

- Cloutier, S., McCallum, B. D., Loutre, C., Banks, T. W., Wicker, T., Feuillet, C., ... Jordan, M. C. (2007). Leaf rust resistance gene *Lr1*, isolated from bread wheat (*Triticum aestivum* L.) is a member of the large psr567 gene family. *Plant Molecular Biology*, *65*, 93–106.
- Collins, N. C., Thordal-Christensen, H., Lipka, V., Bau, S., et al. (2003). SNARE-protein-mediated disease resistance at the plant cell wall. *Nature*, *425*, 973–977.
- Conti, V., Roncallo, P. F., Beaufort, V., Cervigni, G. L., Miranda, R., Jensen, C. A., & Echenique, V. C. (2011). Mapping of main and epistatic effect QTLs associated to grain protein and gluten strength using a RIL population of durum wheat. *Journal of Applied Genetics*, *52*, 287–298.
- Cossani, C. M., & Reynolds, M. P. (2012). Physiological traits for improving heat tolerance in wheat. *Plant Physiology*, *160*, 1710–1718.
- Crespo-Herrera, L., Velu, G., & Singh, R. (2016). Quantitative trait loci mapping reveals pleiotropic effect for grain iron and zinc concentrations in wheat. *The Annals of Applied Biology*, *169*, 27–35.
- Dakouri, A., McCallum, B. D., Radovanovic, N., & Cloutier, S. (2013). Molecular and phenotypic characterization of seedling and adult plant leaf rust resistance in a world wheat collection. *Molecular Breeding*, *32*, 663–677.
- Daryanto, S., Wang, L., & Jacinthe, P. A. (2016). Global synthesis of drought effects on maize and wheat production. *PLoS One*, *11*, e0156362.
- Devos, K. M., & Gale, M. D. (2000). Genome relationships: The grass model in current research. *The Plant Cell*, *12*, 637–646.
- Distelfeld, A., Cakmak, I., Peleg, Z., Ozturk, L., Yazici, A. M., Budak, H., ... Fahima, T. (2007). Multiple QTL-effects of wheat Gpc-B1 locus on grain protein and micronutrient concentrations. *Physiologia Plantarum*, *129*, 635–643.
- Draper, J., Mur, L. J., Jenkins, G., Ghosh-Biswas, C., Bablak, P., Hasterok, R., & Routledge, A. P. M. (2001). *Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant Physiology*, *127*, 1539–1555.
- Durmaz, E., Coruh, C., Dinler, G., Grusak, M. A., Peleg, Z., Saranga, Y., ... Cakmak, I. (2011). Expression and cellular localization of ZIP1 transporter under zinc deficiency in wild emmer wheat. *Plant Molecular Biology Reporter/ISPMB*, *29*, 582–596.
- El Hassouni, K., Alahmad, S., Belkadi, B., Filali-Maltouf, A., Hickey, L., & Bassi, F. (2018). Root system architecture and its association with yield under different water regimes in durum wheat. *Crop Science*, *58*, 2331–2346.
- Ellis, J. G., Lagudah, E. S., Spielmeier, W., & Dodds, P. N. (2014). The past, present and future of breeding rust resistant wheat. *Frontiers in Plant Science*, *5*, 641.
- Evens, N. P., Buchner, P., Williams, L. E., & Hawkesford, M. J. (2017). The role of ZIP transporters and group F bZIP transcription factors in the Zn-deficiency response of wheat (*Triticum aestivum*). *The Plant Journal: For Cell and Molecular Biology*, *92*, 291–304.
- Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A., & Keller, B. (2003). Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proceedings of the National Academy of Sciences*, *100*, 15253–15258.
- Figueroa, M., Hammond-Kosack, K. E., & Solomon, P. S. (2018). A review of wheat diseases—A field perspective. *Molecular Plant Pathology*, *19*, 1523–1536.
- Friebe, B., Heun, M., Tuleen, N., Zeller, F., & Gill, B. (1994). Cytogenetically monitored transfer of powdery mildew resistance from rye into wheat. *Crop Science*, *34*, 621–625.
- Gahlaut, V., Jaiswal, V., Singh, S., Balyan, H., & Gupta, P. (2019). Multi-locus genome wide association mapping for yield and its contributing traits in hexaploid wheat under different water regimes. *Scientific Reports*, *9*, 1–15.
- Gale, M. D., & Devos, K. M. (1998). Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 1971–1974.
- Gallavotti, A., Zhao, Q., Kyozuka, J., Meeley, R. B., Ritter, M. K., Doebley, J. F., ... Schmidt, R. J. (2004). The role of barren stalk1 in the architecture of maize. *Nature*, *432*, 630–635.
- Gao, P., Zhou, Y., Gebrewahid, T. W., Zhang, P., Yan, X., Li, X., ... Liu, D. (2019). Identification of known leaf rust resistance genes in common wheat cultivars from Sichuan province in China. *Crop Protection (Guildford, Surrey)*, *115*, 122–129.
- Garg, M., Chawla, M., Chunduri, V., Kumar, R., Sharma, S., Sharma, N. K., ... Saini, M. K. (2016). Transfer of grain colors to elite wheat cultivars and their characterization. *Journal of Cereal Science*, *71*, 138–144.
- Gaut, B. S., & Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 6809–6814.
- Gautam, T., Saripalli, G., Kumar, A., Gahlaut, V., Gadekar, D., Oak, M., ... Gupta, P. (2020). Introgression of a drought insensitive grain yield QTL for improvement of four Indian bread wheat cultivars using marker assisted breeding without background selection. *Journal of Plant Biochemistry and Biotechnology*, *30*, 172–183.
- Genc, Y., Verbyla, A., Torun, A., Cakmak, I., Willmore, K., Wallwork, H., & McDonald, G. (2009). Quantitative trait loci analysis of zinc efficiency and grain zinc concentration in wheat using whole genome average interval mapping. *Plant and Soil*, *314*, 49.
- Gottwald, S., Stein, N., Borner, A., Sasaki, T., & Graner, A. (2004). The gibberellic-acid insensitive dwarfing gene *sdw3* of barley is located on chromosome 2HS in a region that shows high colinearity with rice chromosome 7L. *Molecular Genetics and Genomics: MGG*, *271*, 426–436.
- Gou, J.-Y., Li, K., Wu, K., Wang, X., Lin, H., Cantu, D., ... Inoue, K. (2015). Wheat stripe rust resistance protein WKS1 reduces the ability of the thylakoid-associated ascorbate peroxidase to detoxify reactive oxygen species. *The Plant Cell*, *27*, 1755–1770.
- Griffiths, S., Dunford, R. P., Coupland, G., & Laurie, D. A. (2003). The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis. *Plant Physiology*, *131*, 1855–1867.
- Griffiths, S., Sharp, R., Foote, T. N., Bertin, I., Wanous, M., Reader, S., ... Moore, G. (2006). Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature*, *439*, 749–752.
- Guan, J., Diego, F., Yun Zhou, G., Appels, R., Li, A., & Mao, L. (2020). The battle to sequence the bread wheat genome: A tale of the three kingdoms. *Genomics, Proteomics and Bioinformatics*, *18*, 221–229.

- Gulick, P. J., Drouin, S., Yu, Z., Danyluk, J., Poisson, G., Monroy, A. F., & Sarhan, F. (2005). Transcriptome comparison of winter and spring wheat responding to low temperature. *Genome National Research Council Canada*, 48, 913–923.
- Gupta, A., Rico-Medina, A., & Caño-Delgado, A. I. (2020). The physiology of plant responses to drought. *Science (New York, N.Y.)*, 368, 266–269.
- Gupta, P. K., Balyan, H. S., Sharma, S., & Kumar, R. (2020). Genetics of yield, abiotic stress tolerance and biofortification in wheat (*Triticum aestivum* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 133, 1569–1602.
- Gupta, O. P., Pandey, V., Gopalareddy, K., Sharma, P., & Singh, G. P. (2019). Genomic intervention in wheat improvement. In S. M. P. Khurana, & R. K. Gaur (Eds.), *Plant biotechnology: Progress in genomics era* (pp. 77–90). Singapore: Springer Nature.
- Gupta, P. K., Langridge, P., & Mir, R. R. (2010). Marker-assisted wheat breeding: Present status and future possibilities. *Molecular Breeding*, 26, 145–161.
- Guyot, R., Yahiaoui, N., Feuillet, C., & Keller, B. (2004). In silico comparative analysis reveals a mosaic conservation of genes within a novel colinear region in wheat chromosome 1AS and rice chromosome 5S. *Functional & Integrative Genomics*, 4, 47–58.
- Hameed, A., Bibi, N., Akhter, J., & Iqbal, N. (2011). Differential changes in antioxidants, proteases, and lipid peroxidation in flag leaves of wheat genotypes under different levels of water deficit conditions. *Plant Physiology and Biochemistry: PPB/Societe Francaise de Physiologie Vegetale*, 49, 178–185.
- Hao, Y., Velu, G., Peña, R. J., Singh, S., & Singh, R. P. (2014). Genetic loci associated with high grain zinc concentration and pleiotropic effect on kernel weight in wheat (*Triticum aestivum* L.). *Molecular Breeding*, 34, 1893–1902.
- Hazratkulova, S., Sharma, R. C., Alikulov, S., Islomov, S., Yuldashev, T., Ziyaev, Z., ... Turok, J. (2012). Analysis of genotypic variation for normalized difference vegetation index and its relationship with grain yield in winter wheat under terminal heat stress. *Plant Breeding*, 131, 716–721.
- He, H., Zhu, S., Zhao, R., Jiang, Z., Ji, Y., Ji, J., ... Bie, T. (2018). *Pm21*, encoding a typical CC-NBS-LRR protein, confers broad-spectrum resistance to wheat powdery mildew disease. *Molecular Plant*, 11, 879–882.
- Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6, 95–108.
- Huang, L., Brooks, S. A., Li, W., Fellers, J. P., Trick, H. N., & Gill, B. S. (2003). Map-based cloning of leaf rust resistance gene *Lr21* from the large and polyploid genome of bread wheat. *Genetics*, 164, 655–664.
- Hurni, S., Brunner, S., Buchmann, G., Herren, G., Jordan, T., Krukowski, P., ... Keller, B. (2013). Rye *Pm8* and wheat *Pm3* are orthologous genes and show evolutionary conservation of resistance function against powdery mildew. *The Plant Journal: For Cell and Molecular Biology*, 76, 957–969.
- IWGSC. (2014). International Wheat Genome Sequencing Consortium A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, 345, 1251788.
- IWGSC. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (New York, N.Y.)*, 361, eaar7191. Available from <https://doi.org/10.1126/science.aar7191>.
- Jaiswal, V., Gahlaut, V., Meher, P. K., Mir, R. R., Jaiswal, J. P., Rao, A. R., ... Gupta, P. K. (2016). Genome wide single locus single trait, multi-locus and multi-trait association mapping for some important agronomic traits in common wheat (*T. aestivum* L.). *PLoS One*, 11, e0159343.
- Jamil, M., Ali, N., Ali, A., & Mujeeb-Kazi, A. (2020). *Spot blotch in bread wheat: Virulence, resistance, and breeding perspectives. Climate Change and Food Security With Emphasis on Wheat* (pp. 217–228). Elsevier.
- Jia, J., Devos, K., Chao, S., Miller, T., Reader, S., & Gale, M. (1996). RFLP-based maps of the homoeologous group-6 chromosomes of wheat and their application in the tagging of *Pm12*, a powdery mildew resistance gene transferred from *Aegilops speltoides* to wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 92, 559–565.
- John, M., & George, J. J. (2018). Tools for sequence assembly and annotation. In *Proceedings of 10th national science symposium*. Gujarat, India: Christ College, February 11, 2018.
- Joppa, L., Du, C., Hart, G. E., & Hareland, G. A. (1997). Mapping gene (s) for grain protein in tetraploid wheat (*Triticum turgidum* L.) using a population of recombinant inbred chromosome lines. *Crop Science*, 37, 1586–1589.
- Keller, B., & Feuillet, C. (2000). Colinearity and gene density in grass genomes. *Trends in Plant Science*, 5, 246–251.
- Kilian, A., Chen, J., Han, F., Steffenson, B., & Kleinhofs, A. (1997). Towards map-based cloning of the barley stem rust resistance genes *rpg1* and *rpg4* using rice as an intergenomic cloning vehicle. *Plant Molecular Biology*, 35, 187–195.
- Krattinger, S. G., Kang, J., Bräunlich, S., Boni, R., Chauhan, H., Selter, L. L., ... Hensel, G. (2019). Abscisic acid is a substrate of the ABC transporter encoded by the durable wheat disease resistance gene *Lr34*. *The New Phytologist*, 223, 853–866.
- Krattinger, S. G., Lagudah, E. S., Spielmeier, W., Singh, R. P., Huerta-Espino, J., McFadden, H., ... Keller, B. (2009). A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science (New York, N.Y.)*, 323, 1360–1363.
- Krishnappa, G., Singh, A. M., Chaudhary, S., Ahlawat, A. K., Singh, S. K., Shukla, R. B., ... Solanki, I. S. (2017). Molecular mapping of the grain iron and zinc concentration, protein content and thousand kernel weight in wheat (*Triticum aestivum* L.). *PLoS One*, 12, e0174972.
- Kulwal, P., Kumar, N., Kumar, A., Gupta, R. K., Balyan, H. S., & Gupta, P. K. (2005). Gene networks in hexaploid wheat: Interacting quantitative trait loci for grain protein content. *Functional & Integrative Genomics*, 5, 254–259.
- Kumar, A., Saripalli, G., Jan, I., Kumar, K., Sharma, P., Balyan, H., & Gupta, P. (2020). *Meta-QTL analysis and identification of candidate genes for drought tolerance in bread wheat (Triticum aestivum L.)*. *Physiology and Molecular Biology of Plants*, 26, 1713–1725.
- Kumar, J., Saripalli, G., Gahlaut, V., Goel, N., Meher, P. K., Mishra, K. K., ... Sansaloni, C. (2018). Genetics of Fe, Zn, β -carotene, GPC and yield traits in bread wheat (*Triticum aestivum* L.) using multi-locus and multi-traits GWAS. *Euphytica*, 214, 219.
- Kumar, S., Röder, M. S., Tripathi, S. B., Kumar, S., Chand, R., Joshi, A. K., & Kumar, U. (2015). Mendelization and fine mapping of a bread wheat spot blotch disease resistance QTL. *Molecular Breeding*, 35, 218.

- Kumar, U., Joshi, A., Kumar, S., Chand, R., & Röder, M. (2010). Quantitative trait loci for resistance to spot blotch caused by *Bipolaris sorokiniana* in wheat (*T. aestivum* L.) lines 'Ning 8201' and 'Chirya 3'. *Molecular Breeding*, 26, 477–491.
- Kumari, M., Pudake, R., Singh, V., & Joshi, A. K. (2013). Association of staygreen trait with canopy temperature depression and yield traits under terminal heat stress in wheat (*Triticum aestivum* L.). *Euphytica*, 190, 87–97.
- Larsen, P. A., Heilman, A. M., & Yoder, A. D. (2014). The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. *BMC Genomics*, 15, 720.
- Leister, D., Kurth, J., Laurie, D. A., Yano, M., Sasaki, T., Devos, K., ... Schulze-Lefert, P. (1998). Rapid reorganization of resistance gene homologues in cereal genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 370–375.
- Li, H., Dong, Z., Ma, C., Xia, Q., Tian, X., Sehgal, S., ... Liu, W. (2020). A spontaneous wheat-*Aegilops longissima* translocation carrying *Pm66* confers resistance to powdery mildew. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 133, 1149–1159.
- Li, J., Dundas, I., Dong, C., Li, G., Trethowan, R., Yang, Z., ... Zhang, P. (2020). Identification and characterization of a new stripe rust resistance gene *Yr83* on rye chromosome 6R in wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 133, 1095–1107.
- Li, Y., Shi, X., Hu, J., Wu, P., Qiu, D., Qu, Y., ... Yang, L. (2020). Identification of a recessive gene *PmQ* conferring resistance to powdery mildew in wheat landrace Qingxinmai using BSR-Seq analysis. *Plant Disease*, 104, 743–751.
- Liang, Z., Chen, K., Li, T., Zhang, Y., Wang, Y., Zhao, Q., ... Ran, Y. (2017). Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. *Nature Communications*, 8, 1–5.
- Lillemo, M., Joshi, A. K., Prasad, R., Chand, R., & Singh, R. P. (2013). QTL for spot blotch resistance in bread wheat line Saar co-locate to the biotrophic disease resistance loci *Lr34* and *Lr46*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 126, 711–719.
- Lopes, M. S., & Reynolds, M. P. (2010). Partitioning of assimilates to deeper roots is associated with cooler canopies and increased yield under drought in wheat. *Functional Plant Biology: FPB*, 37, 147–156.
- Lu, P., Liang, Y., Li, D., Wang, Z., Li, W., Wang, G., ... Xie, J. (2016). Fine genetic mapping of spot blotch resistance gene *Sb3* in wheat (*Triticum aestivum*). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 129, 577–589.
- Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, 1, 18.
- Mains, E. B., Leighty, C., & Johnston, C. (1926). Inheritance of resistance to leaf rust, *Puccinia triticina* Erikss, in crosses of common wheat, *Triticum vulgare* vill. Authority of the Secretary of Agriculture.
- Malviya, N., Yadav, S., & Yadav, D. (2019). Bioinformatics intervention in plant biotechnology: An overview. In S. M. P. Khurana, & R. K. Gaur (Eds.), *Plant biotechnology: Progress in genomics era* (pp. 175–188). Singapore: Springer Nature.
- Manschadi, A. M., Hammer, G. L., Christopher, J. T., & Devoil, P. (2008). Genotypic variation in seedling root architectural traits and implications for drought adaptation in wheat (*Triticum aestivum* L.). *Plant and Soil*, 303, 115–129.
- Mason, R. E., Mondal, S., Beecher, F. W., & Hays, D. B. (2011). Genetic loci linking improved heat tolerance in wheat (*Triticum aestivum* L.) to lower leaf and spike temperatures under controlled conditions. *Euphytica*, 180, 181–194.
- Mathur, S., Agrawal, D., & Jajoo, A. (2014). Photosynthesis: Response to high temperature stress. *Journal of Photochemistry and Photobiology B, Biology*, 137, 116–126.
- Maulana, F., Huang, W., Anderson, J. D., & Ma, X.-F. (2020). Genome-wide association mapping of seedling drought tolerance in winter wheat. *Frontiers in Plant Science*, 11, 573786.
- McDonald, M. C., Ahren, D., Simpfendorfer, S., Milgate, A., & Solomon, P. S. (2018). The discovery of the virulence gene *ToxA* in the wheat and barley pathogen *Bipolaris sorokiniana*. *Molecular Plant Pathology*, 19, 432–439.
- McFadden, E. S. (1930). A successful transfer of emmer characters to vulgare wheat 1. *Agronomy Journal*, 22, 1020–1034.
- McIntosh, R. A., Wellings, C. R., & Park, R. F. (1995). *Wheat rusts: An atlas of resistance genes*. CSIRO Publishing.
- Merchuk-Ovnat, L., Barak, V., Fahima, T., Ordon, F., Lidzbarsky, G. A., Krugman, T., & Saranga, Y. (2016). Ancestral QTL alleles from wild emmer wheat improve drought resistance and productivity in modern wheat cultivars. *Frontiers in Plant Science*, 7, 452.
- Miller, T., Reader, S., Ainsworth, C., & Summers, R. (1998). The introduction of a major gene for resistance to powdery mildew of wheat, *Erysiphe graminis* f. sp. tritici, from *Aegilops speltoides* into wheat, *Triticum aestivum*. In *Conference of the Cereal Section of EUCARPIA (European Association for Research on Plant Breeding)*. Wageningen, Netherlands, 24–26 February 1988. Pudoc.
- Miranda, L. M., Murphy, J. P., Marshall, D., & Leath, S. (2006). *Pm34*: A new powdery mildew resistance gene transferred from *Aegilops tauschii* Coss. to common wheat (*Triticum aestivum* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 113, 1497–1504.
- Misale, C., Ferrero, G., Torquati, M., & Aldinucci, M. (2014). Sequence alignment tools: One parallel pattern to rule them all? *BioMed Research International*, 2014, 539410. Available from <https://doi.org/10.1155/2014/539410>.
- Mishra, S., Singh, S., Patil, R., Bhusal, N., Malik, A., Sareen, S., ... Gupta, R. (2014). Breeding for heat tolerance in Wheat. *Genetics*, 2, 1.
- Monna, L., Kitazawa, N., Yoshino, R., Suzuki, J., Masuda, H., Maehara, Y., ... Minobe, Y. (2002). Positional cloning of rice semidwarfing gene, *sd-1*: Rice 'green revolution gene' encodes a mutant enzyme involved in gibberellin synthesis. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 9, 11–17.
- Moore, J. W., Herrera-Foessel, S., Lan, C., Schnippenkoetter, W., Ayliffe, M., Huerta-Espino, J., ... Periyannan, S. (2015). A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. *Nature Genetics*, 47, 1494–1498.
- Morgounov, A., Gómez-Becerra, H. F., Abugalieva, A., Dzhunusova, M., Yessimbekova, M., Muminjanov, H., ... Cakmak, I. (2007). Iron and zinc grain density in common wheat grown in Central Asia. *Euphytica*, 155, 193–203.

- Mujeeb-Kazi, A., Ali, N., Ibrahim, A., Napar, A. A., Jamil, M., Hussain, S., . . . Rajaram, S. (2017). Tissue culture mediated allelic diversification and genomic enrichment of wheat to combat production constraints and address food security. *Plant Tissue Culture Biotechnology*, 27, 89–140.
- Mujeeb-Kazi, A., Gul, A., Ahmad, I., Farooq, M., Rizwan, S., Bux, H., . . . Delgado, R. (2007). *Aegilops tauschii*, as a spot blotch (*Cochliobolus sativus*) resistance source for bread wheat improvement. *Pakistan Journal of Botany*, 39, 1207.
- Mullan, D. J., & Reynolds, M. P. (2010). Quantifying genetic effects of ground cover on soil water evaporation using digital imaging. *Functional Plant Biology: FPB*, 37, 703–712.
- Navathe, S., Yadav, P. S., Chand, R., Mishra, V. K., Vasistha, N. K., Meher, P. K., . . . Gupta, P. K. (2020). ToxA–Tsn1 Interaction for spot blotch susceptibility in Indian wheat: An example of inverse gene-for-gene relationship. *Plant Disease*, 104, 71–81.
- Ni, Z., Li, H., Zhao, Y., Peng, H., Hu, Z., Xin, M., & Sun, Q. (2018). Genetic improvement of heat tolerance in wheat: Recent progress in understanding the underlying molecular mechanisms. *Crop Journal*, 6, 32–41.
- Nio, S., Cawthray, G., Wade, L., & Colmer, T. (2011). Pattern of solutes accumulated during leaf osmotic adjustment as related to duration of water deficit for wheat at the reproductive stage. *Plant Physiology and Biochemistry: PPB/Societe Francaise de Physiologie Vegetale*, 49, 1126–1137.
- Pandey, R., Joshi, G., Bhardwaj, A. R., Agarwal, M., & Katiyar-Agarwal, S. (2014). A comprehensive genome-wide study on tissue-specific and abiotic stress-specific miRNAs in *Triticum aestivum*. *PLoS One*, 9, e95800.
- Paterson, A. H., Lin, Y. R., Li, Z. K., Schertz, K. F., Doebley, J. F., Pinson, S. R. M., . . . Irvine, J. E. (1995). Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science (New York, N.Y.)*, 269, 1714–1718.
- Peleg, Z., Cakmak, I., Ozturk, L., Yazici, A., Jun, Y., Budak, H., . . . Saranga, Y. (2009). Quantitative trait loci conferring grain mineral nutrient concentrations in durum wheat × wild emmer wheat RIL population. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 119, 353–369.
- Peng, J., Richards, D. E., Hartley, N. M., Murphy, G. P., Devos, K. M., Flintham, J. E., Beales, J., Fish, L. J., Worland, A. J., Pelica, F., et al. (1990). ‘Green revolution’ genes encode mutant gibberellin response modulators. *Nature*, 400, 256–261.
- Periyannan, S., Milne, R. J., Figueroa, M., Lagudah, E. S., & Dodds, P. N. (2017). An overview of genetic rust resistance: From broad to specific mechanisms. *PLoS Pathogens*, 13, e1006380.
- Pinto, R. S., Reynolds, M. P., Mathews, K. L., McIntyre, C. L., Olivares-Villegas, J.-J., & Chapman, S. C. (2010). Heat and drought adaptive QTL in a wheat population designed to minimize confounding agronomic effects. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 121, 1001–1021.
- Prasad, P. V., Pisipati, S., Ristic, Z., Bukovnik, U., & Fritz, A. (2008). Impact of night time temperature on physiology and growth of spring wheat. *Crop Science*, 48, 2372–2380.
- Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology*, 5, 94–100.
- Ram, S., Verma, A., & Sharma, S. (2010). Large variability exists in phytase levels among Indian wheat varieties and synthetic hexaploids. *Journal of Cereal Science*, 52, 486–490.
- Rasheed, A., Mujeeb-Kazi, A., Ogbonnaya, F. C., He, Z., & Rajaram, S. (2018). Wheat genetic resources in the post-genomics era: Promise and challenges. *Annals of Botany*, 121, 603–616.
- Ray, D. K., Mueller, N. D., West, P. C., & Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLoS One*, 8, e66428.
- Rebetzke, G., Appels, R., Morrison, A., Richards, R., McDonald, G., Ellis, M., . . . Bonnett, D. (2001). Quantitative trait loci on chromosome 4B for coleoptile length and early vigour in wheat (*Triticum aestivum* L.). *Australian Journal of Agricultural Research*, 52, 1221–1234.
- Reynolds, M. (1997). *Using canopy temperature depression to select for yield potential of wheat in heat-stressed environments* (Vol. 42). CIMMYT.
- Richard, C. A., Hickey, L. T., Fletcher, S., Jennings, R., Chenu, K., & Christopher, J. T. (2015). High-throughput phenotyping of seminal root traits in wheat. *Plant Methods*, 11, 13.
- Riley, R., Chapman, V., & Johnson, R. (1968a). The incorporation of alien disease resistance in wheat by genetic interference with the regulation of meiotic chromosome synapsis. *Genetical Research*, 12, 199–219.
- Riley, R., Chapman, V., & Johnson, R. (1968b). Introduction of yellow rust resistance of *Aegilops comosa* into wheat by genetically induced homoeologous recombination. *Nature*, 217, 383–384.
- Ruqiang, X., Qixin, S., & Shuzhen, Z. (1996). Chromosomal location of genes for heat tolerance as measured by membrane thermostability of common wheat cv. Hope. *Yi Chuan = Hereditas*, 18, 1–3.
- Salse, J., Piegue, B., Cooke, R., & Delseny, M. (2002). Synteny between *Arabidopsis thaliana* and rice at the genome level: A tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Research*, 30, 2316–2328.
- Sánchez-Martín, J., Steuernagel, B., Ghosh, S., Herren, G., Huml, S., Adamski, N., . . . Wicker, T. (2016). Rapid gene isolation in barley and wheat by mutant chromosome sequencing. *Genome Biology*, 17, 221.
- SanMiguel, P. J., Ramakrishna, W., Bennetzen, J. L., Busso, C. S., & Dubcovsky, J. (2002). Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). *Functional & Integrative Genomics*, 2, 70–80.
- Sarieva, G., Kenzhebaeva, S., & Lichtenhaler, H. (2010). Adaptation potential of photosynthesis in wheat cultivars with a capability of leaf rolling under high temperature conditions. *Russian Journal of Plant Physiology*, 57, 28–36.
- Schadt-Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, 11, 31–46.
- Schwessinger, B. (2017). Fundamental wheat stripe rust research in the 21st century. *The New Phytologist*, 213, 1625–1631.
- Sears, E. (1956). *The transfer of leaf-rust resistance from Aegilops umbellulata to wheat*. *Genetics in plant breeding* (pp. 1–22). Brook-Haven Symposia in Biology.

- Sehgal, D., Autrique, E., Singh, R., Ellis, M., Singh, S., & Dreisigacker, S. (2017). Identification of genomic regions for grain yield and yield stability and their epistatic interactions. *Scientific Reports*, 7, 41578.
- Sharma, P., Sareen, S., Saini, M., Verma, A., Tyagi, B. S., & Sharma, I. (2014). Assessing genetic variation for heat tolerance in synthetic wheat lines using phenotypic data and molecular markers. *Australian Journal of Crop Science*, 8, 515.
- Sharma, S., Xu, S., Ehdai, B., Hoops, A., Close, T. J., Lukaszewski, A. J., & Waines, J. G. (2011). Dissection of QTL effects for root traits using a chromosome arm-specific mapping population in bread wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 122, 759–769.
- Sheikh, I., Sharma, P., Verma, S. K., Kumar, S., Kumar, N., Kumar, S., . . . Dhaliwal, H. S. (2018). Development of intron targeted amplified polymorphic markers of metal homeostasis genes for monitoring their introgression from *Aegilops* species to wheat. *Molecular Breeding*, 38, 47.
- Sheikh, I., Vyas, P., & Dhaliwal, H. S. (2020). *Biofortification of wheat through wide hybridization and molecular breeding. Wheat and barley grain biofortification* (pp. 139–181). Elsevier.
- Shewry, P. R. (2009). Wheat. *Journal of Experimental Botany*, 60, 1537–1553.
- Shi, R., Li, H., Tong, Y., Jing, R., Zhang, F., & Zou, C. (2008). Identification of quantitative trait locus of zinc and phosphorus density in wheat (*Triticum aestivum* L.) grain. *Plant and Soil*, 306, 95–104.
- Shiferaw, B., Smale, M., Braun, H.-J., Duveiller, E., Reynolds, M., & Muricho, M. (2013). Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Security*, 5, 291–317.
- Shubing, L., & Honggang, W. (2005). Characterization of a wheat–*Thinopyron intermedium* substitution line with resistance to powdery mildew. *Euphytica*, 143, 229–233.
- Singh, D., Kumar, D., Satapathy, L., Pathak, J., Chandra, S., Riaz, A., et al. (2017). Insights of *Lr28* mediated wheat leaf rust resistance: Transcriptomic approach. *Gene*, 637, 72–89.
- Singh, R., Govindan, V., & Andersson, M. S. (2017). Zinc-biofortified wheat: Harnessing genetic diversity for improved nutritional quality. *Science Brief: Biofortification*, 1.
- Singh, R., Nelson, J., & Sorrells, M. (2000). Mapping Yr28 and other genes for resistance to stripe rust in wheat. *Crop Science*, 40, 1148–1155.
- Sivamani, E., Bahieldin, A., Wraith, J. M., Al-Niemi, T., Dyer, W. E., Ho, T.-H. D., & Qu, R. (2000). Improved biomass productivity and water use efficiency under water deficit conditions in transgenic wheat constitutively expressing the barley HVA1 gene. *Plant Science (Shannon, Ireland)*, 155, 1–9.
- Song, R., Llaca, V., & Messing, J. (2002). Mosaic organization of orthologous sequences in grass genomes. *Genome Research*, 12, 1549–1555.
- Spielmeyer, W., Hyles, J., Joaquim, P., Azanza, F., Bonnett, D., Ellis, M., . . . Richards, R. (2007). QTL on chromosome 6A in bread wheat (*Triticum aestivum*) is associated with longer coleoptiles, greater seedling vigour and final plant height. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 115, 59–66.
- Srinivasa, J., Arun, B., Mishra, V. K., Singh, G. P., Velu, G., Babu, R., . . . Joshi, A. K. (2014). Zinc and iron concentration QTL mapped in a *Triticum spelta* × *T. aestivum* cross. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 127, 1643–1651.
- Stein, N., Feuillet, C., Wicker, T., Schlagenhauf, E., & Keller, B. (2000). Subgenome chromosome walking in wheat: A 450-kb physical contig in *Triticum monococcum* L. spans the *Lr10* resistance locus in hexaploid wheat (*Triticum aestivum* L.). *Proceedings of the National Academy of Sciences of the United States of America*, 97, 13436–13441.
- Sun, Q., & Quick, J. (1991). Chromosomal locations of genes for heat tolerance in tetraploid wheat. *Cereal Research Communications*, 19, 431–437.
- Tabbitta, F., Pearce, S., & Barneix, A. J. (2017). Breeding for increased grain protein and micronutrient content in wheat: Ten years of the GPC-B1 gene. *Journal of Cereal Science*, 73, 183–191.
- Tang, C., Xu, Q., Zhao, M., Wang, X., & Kang, Z. (2018). Understanding the lifestyles and pathogenicity mechanisms of obligate biotrophic fungi in wheat: The emerging genomics era. *Crop Journal*, 6, 60–67.
- Tang, S., Hu, Y., Zhong, S., & Luo, P. (2018). The potential role of powdery mildew-resistance Gene Pm40 in Chinese wheat-breeding programs in the Post-Pm21 Era. *Engineering*, 4, 500–506.
- Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding and Biotechnology*, 2, 195–212.
- Tikhonov, A. P., Bennetzen, J. L., & Avramova, Z. V. (2000). Structural domains and matrix attachment regions along colinear chromosomal segments of maize and sorghum. *The Plant Cell*, 12, 249–264.
- Tiwari, C., Wallwork, H., Arun, B., Mishra, V. K., Velu, G., Stangoulis, J., . . . Joshi, A. K. (2016). Molecular mapping of quantitative trait loci for zinc, iron and protein content in the grains of hexaploid wheat. *Euphytica*, 207, 563–570.
- Tiwari, V. K., Rawat, N., Chhuneja, P., Neelam, K., Aggarwal, R., Randhawa, G. S., . . . Singh, K. (2009). Mapping of quantitative trait loci for grain iron and zinc concentration in diploid A genome wheat. *Journal of Heredity*, 100, 771–776.
- Tura, H., Edwards, J., Gahlaut, V., Garcia, M., Sznajder, B., Baumann, U., . . . Balyan, H. S. (2020). QTL analysis and fine mapping of a QTL for yield-related traits in wheat grown in dry and hot environments. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 133, 239–257.
- Turner, A., Beales, J., Faure, S., Dunford, R. P., & Laurie, D. A. (2005). The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science (New York, N.Y.)*, 310, 1031–1034.
- Uauy, C. (2017). Wheat genomics comes of age. *Current Opinion in Plant Biology*, 36, 142–148.
- Uauy, C., Distelfeld, A., Fahima, T., Blechl, A., & Dubcovsky, J. (2006). A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science (New York, N.Y.)*, 314, 1298–1301.
- Vashishth, A., Ram, S., & Beniwal, V. (2017). Variability in phytic acid and phytase levels and utilization of synthetic hexaploids for enhancing phytase levels in bread wheat. *Journal of Wheat Research*, 9, 42–46.

- Velu, G., Singh, R. P., Crespo-Herrera, L., Juliana, P., Dreisigacker, S., Valluru, R., ... Mishra, V. K. (2018). Genetic dissection of grain zinc concentration in spring wheat for mainstreaming biofortification in CIMMYT wheat breeding. *Scientific Reports*, 8, 1–10.
- Vendruscolo, E. C. G., Schuster, I., Pileggi, M., Scapim, C. A., Molinari, H. B. C., Marur, C. J., & Vieira, L. G. E. (2007). Stress-induced synthesis of proline confers tolerance to water deficit in transgenic wheat. *Journal of Plant Physiology*, 164, 1367–1376.
- Verma, V., Foulkes, M., Worland, A., Sylvester-Bradley, R., Caligari, P., & Snape, J. (2004). Mapping quantitative trait loci for flag leaf senescence as a yield determinant in winter wheat under optimal and drought-stressed environments. *Euphytica*, 135, 255–263.
- Vinga, S. (2014). Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 15, 376–389.
- Vinga, S., & Almeida, J. (2013). Alignment-free sequence comparison—A review. *Bioinformatics*, 19, 513–523.
- Vogel, J. P., Gu, Y. Q., Twigg, P., Lazo, G. R., Laudencia-Chinguanco, D., Hayden, D. M., ... Coleman-Derr, D. (2006). EST sequencing and phylogenetic analysis of the model grass *Brachypodium distachyon*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 113, 186–195.
- Vollbrecht, E., Springer, P. S., Goh, L., Buckler, E. S., & Martienssen, R. (2005). Architecture of floral branch systems in maize and related grasses. *Nature*, 436, 1119–1126.
- Wang, W., Simmonds, J., Pan, Q., Davidson, D., He, F., Battal, A., ... Akhunov, E. (2018). Gene editing and mutagenesis reveal inter-cultivar differences and additivity in the contribution of TaGW2 homoeologues to grain size and weight in wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 131, 2463–2475.
- Wasson, A. P., Richards, R., Chatrath, R., Misra, S., Prasad, S. S., Rebetzke, G., ... Watt, M. (2012). Traits and selection strategies to improve root systems and water uptake in water-limited wheat crops. *Journal of Experimental Botany*, 63, 3485–3498.
- White, P. J., & Broadley, M. R. (2009). Biofortification of crops with seven mineral elements often lacking in human diets—iron, zinc, copper, calcium, magnesium, selenium and iodine. *The New Phytologist*, 182, 49–84.
- Xia, H., Wang, L., Qiao, Y., Kong, W., Xue, Y., Wang, Z., ... Sizmur, T. (2020). Elucidating the source–sink relationships of zinc biofortification in wheat grains: A review. *Food and Energy Security*, 9, e243.
- Xu, F., Lagudah, E. S., Moose, S. P., & Riechers, D. E. (2002). Tandemly duplicated safener-induced glutathione S-transferase genes from *Triticum tauschii* contribute to genome- and organ- specific expression in hexaploid wheat. *Plant Physiology*, 130, 362–373.
- Xu, Y., An, D., Liu, D., Zhang, A., Xu, H., & Li, B. (2012b). Molecular mapping of QTLs for grain zinc, iron and protein concentration of wheat across two environments. *Field Crops Research*, 138, 57–62.
- Yahiaoui, N., Srichumpa, P., Dudler, R., & Keller, B. (2004). Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene Pm3b from hexaploid wheat. *The Plant Journal: For Cell and Molecular Biology*, 37, 528–538.
- Yan, J., Xue, W.-T., Yang, R.-Z., Qin, H.-B., Zhao, G., Tzion, F., & Cheng, J.-P. (2018). Quantitative trait loci conferring grain selenium nutrient in durum wheat × wild emmer wheat RIL population. *Czech Journal of Genetics and Plant Breeding*, 54, 52–58.
- Yan, L., Loukoiannov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., SanMiguel, P., ... Dubcovsky, J. (2004). The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science (New York, N.Y.)*, 303, 1640–1644.
- Yan, L., Loukoiannov, A., Tranquilli, G., Helguera, M., Fahima, T., & Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene VRN1. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 6263–6268.
- Yang, J., Sears, R., Gill, B., & Paulsen, G. (2002). Quantitative and molecular characterization of heat tolerance in hexaploid wheat. *Euphytica*, 126, 275, 28.
- Zandipour, M., Hervan, E. M., Azadi, A., Khosroshahli, M., & Etmnan, A. (2020). A QTL hot spot region on chromosome 1B for nine important traits under terminal drought stress conditions in wheat. *Cereal Research Communications*, 48, 17–24.
- Zhang, P., Guo, G., Wu, Q., Chen, Y., Xie, J., Lu, P., ... Wang, R. (2020). Identification and fine mapping of spot blotch (*Bipolaris sorokiniana*) resistance gene Sb4 in wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 133, 2451–2459.
- Zhang, S., Zhang, R., Song, G., Gao, J., Li, W., Han, X., ... Li, G. (2018). Targeted mutagenesis using the *Agrobacterium tumefaciens*-mediated CRISPR-Cas9 system in common wheat. *BMC Plant Biology*, 18, 302.
- Zhao, L., Zhang, K.-P., Liu, B., Deng, Z.-Y., Qu, H.-L., & Tian, J.-C. (2010). A comparison of grain protein content QTLs and flour protein content QTLs across environments in cultivated wheat. *Euphytica*, 174, 325–335.
- Zhou, R., Zhu, Z., Kong, X., Huo, N., Tian, Q., Li, P., ... Jia, J. (2005). Development of wheat near-isogenic lines for powdery mildew resistance. *Theoretical and Applied Genetics*, 110, 640–648.
- Zhu, T., Wu, L., He, H., Song, J., Jia, M., Liu, L., ... Du, W. (2020). Bulk segregant RNA-Seq reveals distinct expression profiling in Chinese wheat cultivar jimai 23 responding to powdery mildew. *Frontiers in Genetics*, 11, 474.
- Zielezinski, A., Vinga, S., Almeida, J., et al. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, 18, 186. Available from <https://doi.org/10.1186/s13059-017-1319-7>.
- Zou, S., Wang, H., Li, Y., Kong, Z., & Tang, D. (2018). The NB-LRR gene *Pm60* confers powdery mildew resistance in wheat. *The New Phytologist*, 218, 298–309.

This page intentionally left blank

Prospects of molecular markers for wheat improvement in postgenomic era

Satish Kumar, Disha Kamboj, Chandra Nath Mishra and Gyanendra Pratap Singh

ICAR-Indian Institute of Wheat and Barley Research Institute, Karnal, Haryana, India

20.1 Introduction

Wheat (*Triticum aestivum* L.) is unusual among recently domesticated species in that it spread to every continent except Antarctica after originating in the Fertile Crescent. High gene plasticity is responsible for the wide range of adaptation from 67°N in Scandinavia to 45°S in Argentina, including elevated regions in the tropics and subtropics (Dubcovsky & Dvorak, 2007). This rapid geographic distribution and exposure of wheat to a variety of environments resulted in distinct gene pools of cultivated wheat based on stature, vernalization requirement, photoperiod reaction, grain consistency, and yield stability (Dubcovsky & Dvorak, 2007; Moose & Mumm, 2008; Worland et al., 1998). Wheat is currently one of the three most important food crops, along with rice and maize (FAO, 2017). Wheat is grown on 200 million hectares around the world, supplying one quarter of the world's overall calorie intake. According to FAO (2017) estimates, the global population will reach 9–10 billion people by 2050, with the majority of people residing in developing countries (Africa and South Asia), where wheat products are the most commonly consumed staple foods. However, due to slow yield improvement of 0.8%–1.0% per year, meeting wheat production requirements at that time would be unlikely. Other major challenges in wheat production include (1) the yield potential while maintaining stability, (2) lowering the cost of increased productivity by reducing the need for water, fertilizers, and other inputs, (3) increasing wheat's ability to grow on marginal lands, (4) lowering greenhouse gas emissions, and (5) continuing to protect wheat from emerging climate change threats (Rasheed & Xia, 2019).

Wheat breeders have taken advantage of “transgressive segregation” in hybridization systems by choosing superior traits that maximize yield in specific conditions. Cultural method of wheat improvement relied on trait selection without understanding the molecular processes of inheritance. The “Green Revolution,” for example, demonstrated how it might often result in massive increases in yield. This development was based on the integration of “slow but efficient” pre-breeding efforts that produced new genetic diversity from *Triticeae* species to protect wheat from abiotic as well as biotic stresses (Mujeeb-Kazi et al., 2013), along with increased yield and nutritional quality (Rasheed, Mujeeb-Kazi, Ogbonnaya, He, & Rajaram, 2018a; Tabbita, Pearce, & Barneix, 2017).

Since the 1980s, a variety of researcher needs, constantly evolving technology, the relevance of crop organisms, DNA sequence databases, genomic abundance of polymorphic traits, and other factors have all influenced the creation of new molecular marker systems in plants, including wheat. Current developments in DNA-based molecular marker technology, genotyping platforms, and reference genome sequence have piqued the interest of functional breeders by making a growing amount of DNA sequence knowledge accessible. These advancements allow breeders to scale up the breeding process and increase precision in the selection of plants carrying favorable genes and/or alleles, as well as their favored combinations. In a single sprint, emerging DNA sequence–based molecular markers will classify a vast number of germplasm for sequence polymorphisms across the entire genome. The key goal of this chapter is to summarize recent advances and development in molecular marker technology, as well as their possible applications (e.g., genomics-assisted breeding) in wheat development. Fig. 20.1 depicts the expansion of wheat molecular marker systems. From the discovery of polymerase chain reaction (PCR) to the reference sequence of the wheat genome, the timeline depicts significant events.

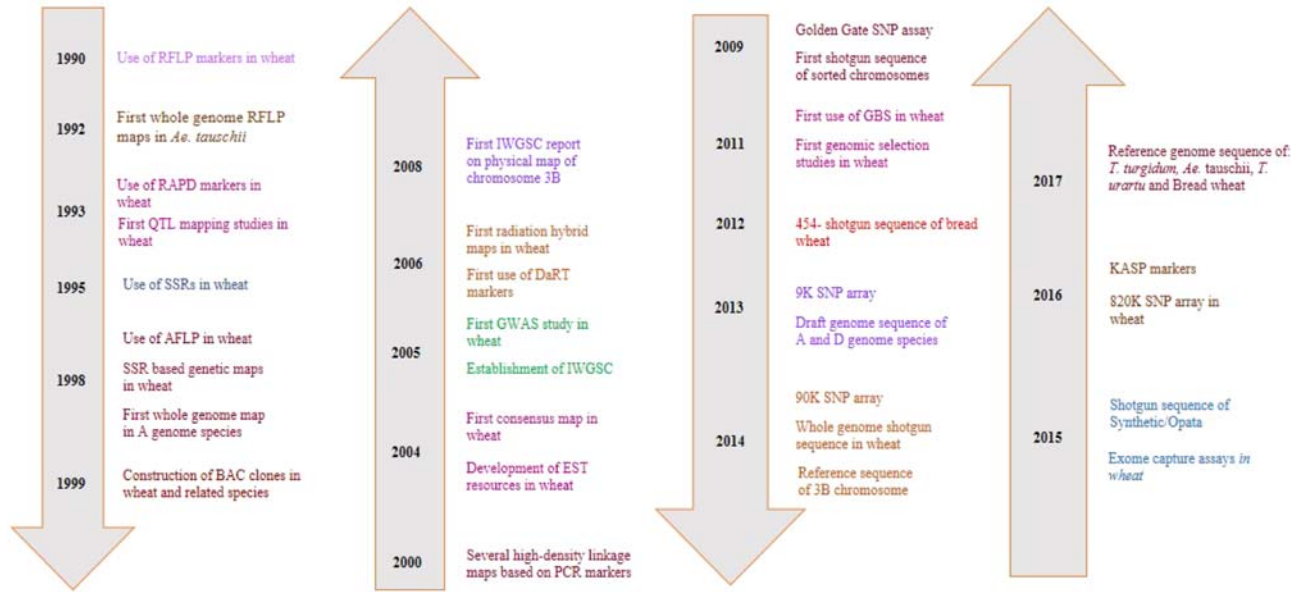


FIGURE 20.1 Growth of molecular marker systems in wheat.

20.2 Overview of molecular marker systems in wheat

Single-nucleotide differences (transitions/transversions), insertions–deletions, and variations in the number and size of tandem repeats at a specific locus are the three most common and special characteristics of the genome. The trick to using such features as a molecular tag to distinguish allelic variation in a gene or detect polymorphism in a single fragment of DNA between two or more individuals is to look at their genome-wide distributions. The positions of these features/tags/molecular markers on the chromosomes can be determined using molecular mapping (Gupta, Varshney, Sharma, & Ramesh, 1999; Landjeva, Korzun, & Borner, 2007). The positions of these features/tags/molecular markers on the chromosomes can be determined using molecular mapping. DNA-based molecular markers are genetic tools that enable plant breeders and geneticists to define and tag genomic regions (QTL (quantitative trait loci) or gene) for particular traits within the genome, and then monitor their inheritance from generation to generation. Since molecular markers have the potential to speed up breeding generations in the field and selection performance in the laboratory, as well as minimize labor and phenotyping costs, these marker systems have been found more effective than traditional plant breeding methods (Langridge & Chalmers, 2004). To fasten the wheat breeding programs, functional markers (gene-specific) have been produced which can differentiate alleles of target genes (Table 20.1). Fingerprinting, trait identification, genome sequencing, genome assembly, comparative mapping, gene cloning, alien gene transfer, and marker-assisted selection (MAS), etc., have been useful in wheat breeding (Table 20.2). Genetic advancement of crop plants would not be feasible without the production and application of molecular markers in agriculture. Furthermore, the pioneering discovery of PCR technology in 1983 revolutionized DNA profiling research and continues to do so today.

The best properties of molecular markers and their uses have been carefully analyzed and discussed (Amom & Nongdam, 2017; Belete, 2018; Gupta et al., 1999; Jiang, 2013; Korzun & Ebmeyer, 2003; Langridge et al., 2001; Prasad, Varshney, Roy, Balyan, & Gupta, 2000; Roder, Huang, & Ganai, 2004; Rustgi, Bandopadhyay, Balyan, & Gupta, 2009; Varshney, Graner, & Sorrells, 2005). If a genetic marker is strongly polymorphic, codominantly hereditary, neutral, evenly distributed across the genome, reproducible, suitable for a variety of applications, and user-friendly, it is called ideal. But in exceptional situations, none of the molecular marker systems will have all of the desired features; based on the scope of analysis, a marker system with all of the necessary features could be favored.

To make the distinction between various marker forms easier, genetic markers are classified into two categories: classical markers and molecular markers. Morphological, physiological, protein/enzyme, and cytological markers are examples of conventional markers. Molecular or DNA sequence–based markers are categorized as (1) hybridization-based, (2) PCR-based, or (3) sequencing-based, depending on their characteristics, production techniques, throughput size, and genotyping/detection procedures.

TABLE 20.1 List of functional markers linked with genes in wheat.

SN	Functional marker	Gene name	References
1	<i>Rht-B1a</i> , <i>Rht-B1b</i> , <i>Rht-D1a</i> and <i>Rht-D1b</i>	<i>Rht-B1</i> and <i>Rht-D1</i>	Ellis, Spielmeier, Gale, Rebetzke, and Richards (2002)
2	<i>vrn-D1</i> , <i>vrn-H1</i> , <i>vrn-B3</i> and <i>vrn-A1</i>	<i>VRN-D1</i> , <i>VRN-H1</i> , <i>VRN-B3</i> and <i>VRN-A1</i>	Fu et al. (2005)
3	<i>PPO18</i>	<i>PPO</i>	Sun et al. (2005)
4	SSR (<i>Pm3a</i> to <i>Pmg</i>)	<i>Pm3</i>	Tommasini, Yahiaoui, Srichumpa, and Keller (2006)
5	<i>Ppd-D1a</i> , <i>Ppd-D1b</i>	<i>Ppd-D1</i>	Beales, Turner, Griffiths, Snape, and Laurie (2007)
6	<i>Ppo-A1a</i> , <i>Ppo-A1b</i>	<i>Ppo-D1</i>	He et al. (2007)
7	<i>YP7B-1</i> , <i>YP7B-2</i> , <i>YP7B-3</i> and <i>YP7B-4</i>	<i>Psy1</i>	He, He, Ma, Appels, and Xia (2009)
8	Five In-Del and one SNP (<i>cssfr1-cssfr6</i>)	<i>Lr34/Yr18/Pm38</i>	Lagudah et al. (2009)
9	SNP	<i>Dreb1</i>	Wei et al. (2009)
10	<i>gluA3a</i> , <i>gluA3b</i> , <i>gluA3d</i> , <i>gluA3e</i> , <i>gluA3f</i> , <i>gluA3g</i> and <i>gluA3ac</i>	<i>Glu-A3</i>	Wang, Li, Peña, Xia, and He (2010)
11	<i>TaGW2-6A</i>	<i>TaGW2-6A</i>	Su, Hao, Wang, Dong, and Zhang (2011)
12	<i>Happa-H</i> and <i>Hap-L</i>	<i>TaSus2-2B</i>	Jiang et al. (2011)
13	<i>TaZds-D1a</i> and <i>TaZds-D1b</i>	<i>TaZds-D1</i>	Zhang et al. (2011)
14	SNP <i>LOX16</i> and <i>LOX8</i>	<i>TaLox-B1</i>	Geng, Xia, Zhang, Qu, and He (2012)
15	<i>TaZds-A1a</i> and <i>TaZds-A1b</i>	<i>TaZds-A1</i>	Dong, Xia, Zhang, and He (2012)
16	SNP	<i>TaMYB2</i>	Garg, Lata, and Prasad (2012)
17	SNP	<i>TaAQP</i>	Pandey, Sharma, Pandey, Sharma, and Chatrath (2013)
18	In-Del	<i>Sr45</i>	Periyannan et al. (2014)
19	<i>POD-3A1</i> and <i>POD-3A2</i>	<i>TaPod-A1</i>	Wei et al. (2015)
20	Two SNPs and one In-Del <i>TaMAMF/TaMAMR</i>	<i>TaMOC1-A</i>	Zhang et al. (2015)
21	SNP (<i>TaGS5-3A-T</i> and <i>TaGS5-3A-G</i>)	<i>TaGS5-3A</i>	Ma et al. (2016)
22	CAPS-SNP	<i>TaTGW6-A1</i>	Hanif et al. (2016)
23	SNP and SSR <i>Xbarc62</i>	<i>TaELF3-1DL</i>	Wang et al. (2016)
24	<i>TaPARM1</i> and <i>TaPARM2</i>	<i>TaPARC</i>	Li et al. (2016)
25	KASP-SNPs (<i>S2269949</i> and <i>S1077313</i>)	<i>CBF-A14</i> under <i>Fr-A2</i> locus	Sieber, Longin, Leiser, and Würschum (2016)
26	<i>TaTPP6AL1-CAPS-F/R</i>	<i>TaTPP-6AL1</i>	Zhang et al. (2017a, b)
27	<i>POD-7D1</i> and <i>POD-7D6</i>	<i>TaPod-D1</i>	Geng, Shi, Fuerst, Wei, and Morris (2019)
28	KASP-SNP	<i>TaSnRK2.9-5A</i>	Rehman et al. (2019)
29	<i>LCY-B1_3765_SNP</i>	<i>TaLcy-B1</i>	Dong et al. (2012)
30	<i>PDS-B1_SNP</i>	<i>Pds-B1</i>	Dong et al. (2012)

TABLE 20.2 List of wheat genes already cloned.

SN	Gene	Linked marker	Method of cloning	References
1.	<i>WX-7A</i> , <i>WX-4A</i> (translocated from 7B), <i>WX-7D</i>	Gene-specific primer design from cDNA sequence AB019622, AB019623 and AB019624	Map-based cloning (In-Del)	Murai, Taira, and Ohta (1999)
2.	<i>Glu-1</i>	<i>Ax1</i> , <i>Ax1</i> , <i>Bx7</i> , <i>Bx17</i> , <i>Dx2</i> , <i>Dx5</i> , <i>By9</i> , <i>Dy10</i> and <i>Dy12</i>	Map-based cloning	De Bustos, Rubio, and Jouve (2001)
3.	<i>pinA</i>	<i>Pina-D1a</i> , <i>b</i> , <i>c</i> , <i>d</i> ; <i>Pinb-D1a</i> , <i>b</i> , <i>e</i> , <i>h</i> , <i>k</i> , <i>l</i> and <i>j</i>	Map-based cloning	Massa, Morris, and Gill (2004) and Guzmán, Caballero, Martín, and Alvarez (2012)
4.	<i>Pinb</i>	<i>Pinb-D1</i>	Map-based cloning	Gautier, Aleman, Guirao, Marion, and Joudrier (1994) and Pan et al. (2004)
5.	<i>B</i>	<i>Xpsr680-7B</i> and <i>Xpsr160-7D</i>	RFLP marker	Jefferies et al. (2000)
6.	<i>Cre3</i>	<i>Xglk605</i> and <i>Xcdo588</i>	RFLP marker	Ogbonnaya et al. (2001)
7.	<i>Rlnn1</i>	<i>Xpsr121</i> , <i>Xpsr680</i> , and <i>Xcdo347</i>	RFLP marker	Williams et al. (2002)
8.	<i>Lr10</i>		Map-based cloning	Feuillet et al. (2003)
9.	<i>Pm</i> , <i>Pm3b</i>	<i>WHS179</i> RFLP marker	Map-based cloning	Yahiaoui, Srichumpa, Dudler, and Keller (2004)
10.	<i>Lr21</i>		Map-based cloning	Huang et al. (2003)
11.	<i>VRN1</i>	<i>WG644</i>	Map-based cloning	Yan et al. (2003)
12.	<i>Nax1</i>	<i>Xgwm312</i> and <i>Xwmc170</i>	Syteny-based cloning	Lindsay, Lagudah, Hare, and Munns (2004)
13.	<i>R genes</i>	<i>Tamyb10-A1</i> , <i>B1</i> , <i>D1</i> (transcription factors)	Expression (transcription factors)	Himi and Noda (2005)
14.	<i>TaNAM</i>	<i>Xuhw106</i> and <i>Xucw109</i>	RNAi expression based	Uauy, Distelfeld, Fahima, Blechl, and Dubcovsky (2006) and Distelfeld et al. (2007)
15.	<i>Lr1</i>	<i>Xpsr567</i>	Map-based cloning	Cloutier et al. (2007)
16.	<i>Psy1</i>	<i>YP7A</i>	Syteny-based cloning	He et al. (2007)
17.	<i>Ppd-D1 (2D)</i>			Beales et al. (2007)
18.	<i>TaVp1</i>	<i>TaVp-A1</i> , <i>TaVp-B1</i> and <i>TaVp-D1</i>	Syteny-based cloning	Utsugi, Nakamura, Noda, and Maekawa (2008)
19.	<i>TaABI5</i>	<i>TaABI5-F/R</i> and <i>qTaABI5-F/R</i>	Map-based cloning	Ohnishi, Himi, Yamasaki, and Noda (2008)
20.	<i>Glu-A1</i> , <i>Glu-D1</i>	<i>UMN19</i> , <i>UMN25</i> and <i>UMN26</i>	Map-based cloning	Liu, Chao, and Anderson (2008)
21.	<i>Lr34/Yr18/ Sr57/ Pm38</i>	<i>Xgwm1220</i> and <i>SWM10</i>	Map-based cloning	Krattinger et al. (2009)
22.	<i>Yr36</i>	<i>Xucw129</i> and <i>Xucw148</i>	Map-based cloning	Fu et al. (2009)

(Continued)

TABLE 20.2 (Continued)

SN	Gene	Linked marker	Method of cloning	References
23.	<i>Utd1</i>	<i>Xgwm234</i> and <i>Xgwm443</i>	Map-based cloning	Randhawa, Popovic, Menzies, Knox, and Fox (2009)
24.	<i>Tsn1</i>	<i>Xfcp623</i>	Map-based cloning	Faris et al. (2010)
25.	<i>DOG-1</i>	<i>DOG1</i> -like genes	Synteny-based cloning	Ashikawa, Abe, and Nakamura (2010)
26.	<i>TaGW2</i>	<i>Xcfd80.2</i>	Synteny-based cloning	Su et al. (2011)
27.	<i>TmMla1</i>	<i>sbi369</i> and <i>sbi314</i>	Synteny-based cloning	Jordan et al. (2011)
28.	<i>Sus2</i>	<i>Xgwm122</i> and <i>Xgwm328</i>	Map-based cloning	Jiang et al. (2011)
29.	<i>TaMFT-3A</i>	<i>CSZENSSR-F1</i> and <i>CSZENSSR-R1</i>	Map-based cloning	Nakamura et al. (2011)
30.	<i>TaCwi-A1</i>	<i>cwi21</i> and <i>cwi22</i>	Synteny-based cloning	Ma, Yan, He, Wu, and Xia (2012)
31.	<i>Pm8</i>	<i>sfr43(Pm8)</i>	Synteny-based cloning	Hurni et al. (2013)
32.	<i>Sr33</i>	<i>BE405778</i> and <i>BE499711</i> (EST markers)	Map-based cloning	Periyannan et al. (2013)
33.	<i>Sr35</i>	<i>AK331487</i> (0.02 cM) and <i>AK332451</i> (0.98 cM)	Map-based cloning	Saintenac et al. (2013a, 2013b)
34.	<i>Yr10</i>	<i>Xpsp3000</i>	Map-based cloning	Liu et al. (2014)
35.	<i>TaSdr-A1</i> , <i>TaSdr-B1</i> , and <i>TaSdr-D1</i>	<i>Sdr-2</i> , <i>Sdr-3</i> and <i>Sdr-4</i>	Map-based cloning	Zhang, Miao, Xia, and He (2014)
36.	<i>Sr50</i>	<i>Sr50-F1/R1</i>	Map-based cloning	Mago et al. (2015)
37.	<i>Lr67/Yr46/ Sr55/ Pm46</i>	<i>Xgwm165</i>	Map-based cloning	Moore et al. (2015)
38.	<i>Snn1</i>	<i>Xfcp618</i> and <i>Xfcp624</i>	Map-based cloning	Shi et al. (2016)
39.	<i>Fhb1</i>	<i>STS3B-355</i> and <i>STS3B-334</i>	Map-based cloning	Rawat et al. (2016)
40.	<i>Phs-A1</i>	<i>Xbarc170</i> and <i>Xwmc420</i>	Map-based cloning	Shorinola et al. (2016)
41.	<i>TaTGW-7A</i>	<i>Xbarc174</i> and <i>Xbarc222</i>	Map-based cloning	Hu et al. (2016)
42.	<i>Sr13</i>	<i>EX24785</i>	Map-based cloning	Zhang et al. (2017a, 2017b)
43.	<i>Stb6</i>	<i>Xctg8311</i> and <i>Xcfn80023</i> (cosegregated with <i>stb6</i>), <i>cfn80025</i> and <i>cfn80030/ cfn80040</i>	Map-based cloning	Saintenac et al. (2018)
44.	<i>Yr15</i>	<i>uhw264</i> and <i>uhw258</i>	Map-based cloning	Klymiuk et al. (2018)

20.3 Genome-wide markers for gene mapping

Continuous advancement in the production of DNA-based markers and genotyping methods has provided useful assistance in the efficient collection of economically important traits over the last three decades (Mir & Varshney, 2013). Since DNA-based markers are plentiful, neutral, stable, simple to automate, and cost-effective, they are favored over conventional markers. Restriction fragment length polymorphisms (RFLPs) were the first hybridization-based markers to be developed and used in wheat for genetic diversity analysis, genetic map creation, and gene tagging (Chao et al., 1989).

In the beginning, RFLP markers were used to establish genetic and physical maps in wheat. In wheat, these efforts culminated in the mapping of over 2000 RFLP loci in genetic maps using segregating populations and over 1200 RFLP loci in physical maps using nullisomic–tetrasomic and deletion lines of Chinese Spring (Gupta et al., 1999; Gupta, Mir, Mohan, & Kumar, 2008a; Gupta, Rustgi, & Mir, 2008b; Hussain & Qamar, 2007). Since DNA probes from one species may easily hybridize with probes from similar species, RFLP markers have proven useful in comparative mapping studies (Devos & Gale, 1993). RFLP markers have been discouraged from further use in wheat genetic studies due to their low-to-medium degree of polymorphism, low-throughput nature of identification, high cost of genotyping, and other factors. Following advancements in genotyping technologies and access to public genomic databases, PCR-based molecular markers such as random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), simple sequence repeat (SSR or microsatellite), diversity arrays technology (DArT), and single-nucleotide polymorphism (SNP) were created.

RFLP and RAPD marker systems were difficult to use in gene mapping/discovery and marker-based selection in the beginning. Following that, RFLP was translated to an AFLP marker method using the PCR technique (Vos et al., 1995). Instead of southern hybridization, PCR amplification was used in AFLP fingerprinting, which allowed for the fractionation of several fragments and the generation of a large number of bands, making polymorphism detection easier. Wheat genetic variation, phylogenetic analysis, and mapping have also been studied using AFLP markers (Bohn, Utz, & Melchinger, 1999; Burkhamer, Lanning, Martens, Martin, & Talbert, 1998; Gupta et al., 1999; Parker, Chalmers, Rathjen, & Langridge, 1999). Similarly, the genomic regions amplified by RAPD markers and linked to variance in targeted traits were cloned, sequenced, and translated into simple, robust, and user-friendly PCR-based markers known as sequence-characterized amplified regions (SCARs).

SSR, DArT, and SNP marker systems for wheat have been established in the past and are now used for a variety of purposes. SSRs can be used in both coding and noncoding areas of the genome and have a wide range of length variation (Bryan et al., 1997; Devos, Moore, & Gale, 1995; Gupta & Varshney, 2000; Roder, Plaschke, König, Börner, & Sorrells, 1995; Zane, Bargelloni, & Patarnello, 2002). SSRs have been found to have much more polymorphism than RAPD, RFLP, and AFLP markers (Bryan et al., 1997; Gupta et al., 2002; Korzun, Roder, Worland, & Börner, 1997; Ma, Roder, & Sorrells, 1996; Plaschke, Ganai, & Roder, 1995; Roder et al., 1995).

There are approximately 3000 and 2000 SSR loci in the available wheat genetic and physical maps prepared using SSR markers, respectively (Goyal et al., 2005; Gupta et al., 2008a, 2008b; Kumar, Goyal, Mohan, Balyan, & Gupta, 2013; Sourdille et al., 2004). Despite the fact that SSRs have become the most common markers for mapping and tagging QTL/genes, their use in wheat genomics has been limited due to (1) the small number of SSR motifs in the genome, (2) their irregular distribution, (3) low-throughput gel-based genotyping, and (4) their inability to multiplex. To address these issues, LGC (<https://www.lgcgroup.com/>) recently launched a new service (<https://www.biosearchtech.com/services/sequencing/microsatellite-ssr-conversion-service>) that uses cutting-edge techniques to transform SSR markers into stable, high-throughput, and cost-effective markers. When opposed to one another, different methods for the production of molecular markers have advantages and drawbacks (Agarwal, Shrivastava, & Padh, 2008; Belete, 2018; Kesawat & Das, 2009), but the reasons for choice vary depending on the needs of consumers.

20.4 Wheat genomics for development of marker and its utilization

Thanks to their abundance in the genomes, versatility for high-throughput genotyping/detection formats, and comparatively low cost, SNP markers have quickly risen to the top of the available molecular markers not only in wheat but also in other crops. SNP markers are the most basic kind of molecular markers, allowing alleles of a gene to differ by a single base pair (nucleotide; DNA building block). The smallest unit of inheritance is a single nucleotide (any of A, T, G, or C) in a gene sequence, and the smallest unit of genetic difference is an SNP.

SNPs are biallelic molecular markers that have variations of four distinct nucleotides. SNPs may be categorized as transformations (A/G or T/C) or transversions (A/T, A/C, G/T, or G/C) based on nucleotide substitution. In a DNA

sequence an SNP could substitute one nucleotide guanine (G) with the nucleotide adenine (A). This SNP variant can be seen in both the coding (exons) and noncoding (introns) parts of a genome, as well as intergenic regions between genes. SNPs are normally discovered *in silico* using preexisting databases with expressed gene tags (ESTs) or sequences from genome surveys (Picoult-Newberg et al., 1999).

SNP markers for the disease-resistant gene Lr34/Yr18/Pm38, which provides resistance to several fungal pathogens, were identified and established by Lagudah, Krattinger, and Herrera-Foessel (2009) in wheat. The production of SNP markers in close proximity to QTL/genes was supported by the genome-wide spread of SNP variants. Following exome sequencing of eight wheat varieties, Allen et al. (2013) reported 10,251 codominant SNPs from 95,266 putative SNPs (Alchemy, Avalon, Cadenza, Hereward, Rialto, Robigus, Savannah, and Xi19). These codominant SNP markers and map can be used to characterize germplasm, conduct QTL experiments, and perform MAS. As a result of these advancements in wheat genomics and breeding, various QTL and underlying genes regulating economically significant traits have been identified, tagged, and cloned (Gale, 2005; Gupta & Varshney, 2000; Landjeva et al., 2007; Lorz & Wenzel, 2004; Nadeem et al., 2018; Mir, Hiremath, Riera-Lizarazu, & Varshney, 2013).

20.5 Status of genotyping platform of bread wheat and its progenitors

The development of robust markers to detect introgressed segments of QTL/gene in the history of recipient wheat genotype is a major challenge for wheat geneticists and breeders. Thanks to their broad and standardized distribution in the genome, SNP markers are currently dominating in genetic research, and their detection relies on the comparison of homologous sequences between genotypes to identify allelic differences at the single-nucleotide stage (Ganal, Altmann, & Roder, 2009; Paux, Sourdille, Mackay, & Feuillet, 2011; Przewieslik-Allen et al., 2019; Rimbart et al., 2018; Wang, Wong, Forrest, & Allen, 2014). Next-generation DNA sequencing (NGS) technologies (such as Illumina's HiSeq, Roche Applied Science's 454, Life Technologies' SOLiD) have greatly increased the discovery of whole-genome SNPs at ever-lower costs (Berkman et al., 2013; Mardis, 2008; for review see Gupta, Rustgi, & Mir, 2013 for review). Because of their potential to uncover a vast number of SNPs from entire genomes, these NGS platforms have captivated consumers (Allen et al., 2011; Bajgain, Rouse, Tsilo, Macharia, & Bhavani, 2016; Cavanagh et al., 2013; Elshire, Glaubitz, & Sun, 2011; Lai et al., 2012; Poland et al., 2012b; Saintenac, Jiang, Wang, & Akhunov, 2013a; Saintenac et al., 2013b; Wang et al., 2014; Winfield et al., 2016). Several technologies for SNP genotyping have been used concurrently with SNP exploration, ranging from low-throughput to high- and ultrahigh-throughput in wheat (Cubizolles et al., 2016; Edwards, Reid, Coghill, Berry, & Barker, 2009; Rimbart et al., 2018). Two relevant advanced technologies [Microarray-Based Genotyping and Genotyping-by-Sequencing (GBS)] for genome-wide SNP discovery and subsequent mapping that are currently being used in wheat are explored further in the chapter.

20.5.1 High-throughput SNP genotyping: microarray-based genotyping

The identification of SNPs from whole-genome and/or transcriptome sequencing using NGS technologies is needed for the development of SNP arrays. This database of DNA/cDNA sequences (NGS reads) is a fantastic tool for detecting SNPs. In addition to NGS-based SNP detection, genomic sequences or EST sequences from various libraries have recently been used for SNP recognition. Clevenger, Chavarro, Pearl, Ozias-Akins, and Jackson (2015) presented an overview of experimental approaches to SNP calling in polyploid organisms such as wheat. Microarrays based on fixed sets of SNP assays have recently been developed by Illumina (Illumina, San Diego, the United States) and Affymetrix (Affymetrix Inc., Santa Clara, CA) for large-scale SNP genotyping (Gupta et al., 2008a, 2008b).

The Infinium II assay system performs whole-genome amplification through a single-base extension step and distinguishes two alleles of a known SNP by incorporating two hapten-labeled dideoxynucleotides (ddNTPs), namely dinitrophenol (red fluorescence) for adenosine (A) and thymine (T) and biotin (blue fluorescence) for cytosine (C) and guanine (G). Infinium II assay uses two fluorescence color assays, so signals have two intensity values per locus depending on allele forms (Gunderson, 2009). The Illumina iScan device scans the fluorescence signals of the assay matrix for more data visualization in the diploid and polyploid versions of the software GenomeStudio. The identification of a significant number of SNPs necessitates high-throughput genotyping. Illumina currently provides a number of custom genotyping array solutions, including the Illumina Infinium iSelect HD chip, which allows unrestricted access to queried SNPs. Wheat has been successfully engineered and used with high-density Infinium arrays for whole-genome SNP genotyping. The International Wheat SNP Working Group (IWSWG) developed the Infinium 9K and 90K iSelect SNP genotyping arrays in partnership with Illumina (Cavanagh et al., 2013; Wang et al., 2014). 7504 SNPs were detected using a 9K iSelect SNP array, and a consensus genetic map of wheat was generated with an average

density of 1.9 1.0 SNP/cM (Cavanagh et al., 2013). A total of 46,977 SNP markers on wheat chromosomes were mapped using the 90K iSelect SNP array (Liu et al., 2016; Wang et al., 2014). Following that, the 90K iSelect SNP array was used for phylogenetic analysis (Turuspekov, Plieske, Ganal, Akhunov, & Abugalieva, 2015), QTL analysis for preharvest sprouting tolerance (Cabral et al., 2014), loose smut resistance (Kumar et al., 2018), leaf rust resistance (Gao et al., 2016a; Gao, Turner, Chao, Kolmer, & Anderson, 2016b; Kumar et al., 2019), physiological traits (Gao et al., 2016a, 2016b) and agronomic traits (Zou et al., 2016); and genome-wide association analysis (GWAS) (Alomari et al., 2019; Garcia et al., 2019; Li, Wen, & Liu, 2019; Liu et al., 2018; Liu, He, & Rasheed, 2017). Gao, Zhao, Huang, and Jia (2017) recently discovered 7989 iSelect SNP loci involved in wheat domestication and improvement, as well as a first-generation map of selection loci for evolutionary studies and breeding. In the genic, repeated, and nonrepetitive intergenic fractions of 8 wheat lines, Rimbart et al. (2018) discovered 3.3 million SNPs a year later. TaBW280K is a high-throughput SNP genotyping array that they created. A biparental population originating from a cross between Chinese Spring and Renan was genotyped using the TaBW280K SNP array, resulting in an ultrahigh-density genetic map of 83,721 SNP markers.

In addition, the Affymetrix Axiom framework has produced a significant number of high-density wheat genotyping SNP arrays. Jordan et al. (2015) previously reported 1.57 million SNPs in 107 Mb sequences from nonredundant low-copy genic regions in 62 wheat genotypes. Winfield et al. (2016) used exome sequencing (exome-seq) to collect 57 Mb of coding sequences in 43 hexaploid wheat accessions and found 9,21,705 (921K) putative SNPs. Of these, 820K high-quality SNPs were used in an array and used to genotype 475 wheat and related accessions. Following that, 35,143 strongly polymorphic and uniformly distributed SNP markers were selected from the 820K SNP sample, and a 35K SNP genotyping array (also known as Wheat Breeder's Array) was built on the Affymetrix GeneTitan platform (Allen et al., 2017). This Wheat Breeder's Array includes SNP markers that were used to identify 2713 wheat genotypes, including landraces, elite lines, and five mapping populations (Allen et al., 2017).

Furthermore, the Wheat660K SNP array, engineered by the Chinese Academy of Agricultural Sciences (<https://wheat.pw.usda.gov/ggpages/topics/Wheat660> SNP array developed by CAAS.pdf) and synthesized by Affymetrix Axiom, has been usable for a wide variety of possible wheat applications. Cui et al. (2017) used an Affymetrix Wheat660K SNP array to genotype 188 recombinant inbred lines (RILs) derived from a cross between KN2904 and J411 to create an ultrahigh-density genetic map of 1,19,566 SNP loci. Using a high-density SNP map and phenotypic results, a big stable QTL (qKnps-4A) for kernel number per spike was discovered (Cui et al., 2017). Comparative genomic research was performed on the genomic sequences of rice (*Oryza sativa*), thale cress (*Brachypodium distachyon*), sorghum (*Sorghum bicolor*), and maize (*Zea mays*) using mapped SNP flanking sequences and corresponding contig sequences of wheat. Furthermore, a new Affymetrix Wheat55K SNP array was created using 53,063 SNP sequence tags carefully chosen from the Wheat660K SNP array. The Wheat55K collection featured SNP tags that were evenly distributed in all 21 wheat chromosomes (2600 SNPs per chromosome) with an average distance of 0.1 cM and a physical distance of approximately 300 kb (Ren et al., 2018).

Cui et al. (2017) and Rimbart et al. (2018) created and communicated two ultrahigh-density genetic maps of SNP markers that can be used to map and dissect complex traits in hexaploid and tetraploid wheat. To date, no publicly available wheat genetic maps have the same SNP density as Cui et al. (2017) and Rimbart et al. (2018). International Wheat Genome Sequencing Consortium (IWGSC) has also used these two SNP maps to anchor and order the wheat genome reference sequence. The studies cited previously show the importance and strength of array-based SNP genotyping in wheat. Due to a variety of advantages such as nucleotide-level variation detection, versatility, speed, and cost-effectiveness, array-based SNP genotyping technologies have gained popularity among users (Thomson, 2014).

Researchers can conduct genetic and physical imaging, marker–trait comparisons, and evolutionary relationship investigations by evaluating existing SNP genotyping platforms. However, owing to the usage of a small collection of wheat germplasm for designing SNP arrays, the genotyping results obtained using the arrays could have determination bias. GBS, a more modern alternative to genotyping technologies, may be used to address these drawbacks. The discovery of a large number of SNPs, as well as the growth of Infinium and Axiom arrays, have provided the wheat community with useful information and tools that could revolutionize wheat breeding.

20.5.2 High-throughput SNP genotyping: genotyping-by-sequencing

NGS for genetic analysis has evolved from a limited number of loci to hundreds of thousands of SNPs due to its growing adaptability and affordability. Prior to sequencing, the reduced-representation sequencing (RRS) method captures only basic DNA regions flanked by restriction enzymes, eliminating genome complexity. There are at least 13 different approaches in the RRS approach family (Scheben, Batley, & Edwards, 2017), 1 of which is GBS. Because of its

simplicity, rapidity, and robustness, the GBS, which was first implemented in maize by [Elshire et al. \(2011\)](#) and later in barley and wheat by [Poland, Brown, Sorrells, and Jannink \(2012a\)](#), is becoming a common genotyping process.

As an alternative to genotyping systems that target a single polymorphism, GBS samples entire DNA for sequencing with certain average sequence depth (depending on the species being used) of genome. GBS primarily depends on restriction endonucleases to catch only the portion of the genome flanked by restriction sites and uses one or two restriction endonucleases to capture only the portion of the genome flanked by restriction sites ([Elshire et al., 2011](#); [He, Holme, & Anthony, 2014](#)). This method of genotyping necessitates the use of high-quality genomic DNA at the required concentration for library planning ([Davey & Blaxter, 2011](#)). GBS has been shown to be effective in predicting breeding values in wheat ([Poland et al., 2012a](#)) and other crop plants by genomic selection and genetic studies ([Bhatia, Wing, & Singh, 2013](#)). GBS provides for the identification of polymorphisms due to the presence/absence of variants in addition to SNPs ([Deschamps, Llaca, & May, 2012](#)). For high-density genetic maps, marker–trait interaction, and genomic selection for days to heading, thousand grain weight, and yield, the GBS for high-throughput genotyping has been commonly used in wheat ([Bhatta, Morgounov, Belamkar, & Baenziger, 2018a](#); [Gao et al., 2017](#); [He et al., 2014](#); [Jamil et al., 2019](#); [Poland & Rife, 2012](#); [Poland et al., 2012a, 2012b](#)). This method has also been used in wheat to map genes/QTLs for disease and insect resistance, as well as preharvest sprouting tolerance ([Bhatta, Morgounov, Belamkar, Yorgancılar, & Baenziger, 2018b](#); [Forrest et al., 2014](#); [Gao et al., 2015](#); [Li et al., 2015a, 2015b](#); [Lin et al., 2015](#); [Zhao et al., 2019](#)). Despite the fact that GBS has the capacity to recognize millions of SNPs, higher levels of missed data (incomplete SNP data) caused by inadequate sequencing coverage often reduce the number of SNPs that can be used for downstream study ([Elshire et al., 2011](#)). In large datasets like GBS, missing data occur when certain experimental lines lack a genotype value at a specific locus, but it is correctly identified and labeled in the remaining lines. The volume of missing data can be reduced by using high-quality genomic DNA, streamlined sequence depth, effective GBS library planning, and sequencing precision.

In wheat and other cereals an improved version of the GBS protocol was developed and used to raise insightful SNPs at a low cost ([Poland et al., 2012a, 2012b](#); [Huang, Poland, Wight, Jackson, & Tinker, 2014](#)). The use of imputation approaches to deal with incomplete data has piqued interest. Genotype imputation is a method of estimating missing genotypes using statistical algorithms such as IMPUTE and fastPHASE; as a result, any value for missing data may be calculated using logical values based on the available reference genome sequence ([Torkamaneh & Belzile, 2015](#)); however, the precision of predicted missing data can be dependent on the reference genome's completeness. [Alipour et al. \(2019\)](#) recently demonstrated how to impute missing genotype data provided by GBS in wheat and barley using the reference genome. The authors found that among the four reference genomes tested (the CSSS, W7984, and IWGSC RefSeq v1.0 wheat reference genomes, and the barley reference genome), IWGSC RefSeq v1.0 imputed the most incomplete SNP data points with sufficient imputation precision. GBS provides a quick, easy, and effective technology of choice for simultaneous detection and genotyping SNPs for genomic-assisted breeding in wheat improvement when combined with data imputation.

20.6 Utility and achievement of high-throughput genotyping approaches in wheat

In a polyploid species like common wheat, which has a massive and complex genome, it has been important for identifying genome-wide SNPs using NGS technologies. It is now possible to search the whole genome for SNP discovery and variation using current rapid sequencing technologies (NGS) and appropriate computer software. The availability of a high-quality reference genome for Chinese Spring wheat has accelerated the resequencing of germplasm accessions and population lines to reliably diagnose SNP variations even within breeding lines that are quite close.

SNP markers are currently showing potential in wheat breeding and genomic science, and they are helping to analyze diverse traits in all modern breeding programs. These markers, for example, have given researchers a better understanding of genetically nuanced traits, including drought and heat resistance. Drought resistance in wheat is governed by a large number of QTLs (or polygenes) with a limited influence. Water-use efficiency, root system architecture, coleoptile length, stomatal conductance, canopy temperature, carbon isotope discrimination, plant phenology, grain yield, and related traits are among the drought-responsive traits ([Ahmad, Ali, & Ahmad, 2017](#); [Gupta, Balyan, & Gahlaut, 2017](#)). Despite the fact that a variety of QTL for the above traits have been detected and mapped, low-density genetic maps have resulted in these QTLs being located at significant intervals between flanking markers. The huge difference between QTLs and flanking markers has made it difficult to use QTLs in wheat breeding by MAS. As a result, high-throughput SNP genotyping approaches (array-based and GBS) have been used to generate a significant number of useful SNP markers that are closely correlated with targeted trait QTL/genes.

A GWAS for yield and related traits under rain-fed conditions was recently conducted on a panel of 123 wheat cultivars from Pakistan (1947–2015) using the Infinium 90K SNP genotyping assay (Ain et al., 2015). This resulted in the discovery of 14,960 polymorphic SNPs, as well as 44 marker–trait associations (MTAs) for 9 yield-related traits. On seven distinct wheat chromosomes, nine multitrait MTAs were identified. Gene annotation of the 44 MTAs, as well as their syntenic relationships to genes in rice, brachypodium, and sorghum, allowed the discovery of 14 MTAs that encode proteins that are expressed in response to stress conditions (Ain et al., 2015). The Seeds of Discovery (Seed) software at CIMMYT used GBS to investigate 1423 spring wheat accessions for a variety of essential traits, including drought and heat resistance (Sehgal et al., 2015). They found 1273 GBS-SNPs in drought-adapted landraces and 4473 SNPs in landraces adapted to heat stress conditions. To make the most of the marker data, more than 200 landraces and synthetic wheat were chosen for their potential use in prebreeding and allele mining of drought and heat stress tolerance candidate genes. Synthetic wheat accessions were found to be more varied than landraces and elite cultivars, according to the mean diversity index. According to the findings, unexplored genetic diversity in landraces and synthetic hexaploid wheat accessions can be characterized and mobilized into well-adapted common cultivars (Sehgal et al., 2015). Although several experiments have been done, only a few have compared genotypic datasets from array- and GBS-based approaches (Elbasyoni et al., 2018; Torkamaneh & Belzile, 2015).

While the SNPs obtained from array-based genotyping are of high quality, the cost per sample is significantly higher. SNP data collected from the GBS network, on the other hand, are greater in volume but have a high proportion of missed calls. The discovery of new SNPs is not possible with array-based genotyping, which is not the case with GBS. Nonetheless, both genotyping technologies are complementary for detecting and mapping essential QTL/genes, based on available SNP genotyping results (Negro, Millet, & Madur, 2019). Elbasyoni et al. (2018) recently compared SNP genotyping data from a 90K SNP array and GBS in winter wheat to estimate population structure and genomic kinship. GBS-scored SNPs are equal to or higher than 90K SNP array-scored SNPs for genomic prediction, according to the authors. The various genotyping technology choices should be carefully analyzed in light of the intended purposes and objectives.

20.7 Conversion of trait-linked SNPs to user-friendly markers

As previously said, the most favored technologies for multiplexing and high-throughput SNP analysis in trait exploration are array-based genotyping and GBS. If a limited number of chosen SNPs are to be genotyped on a huge set of germplasm and breeding lines, they are inflexible and costly. It is critical to find an SNP assay that's modular, cost-effective, user-friendly, and time-saving, as well as generates high-quality results. LGC Genomics (<http://www.lgcgroup.com/>) solved this research challenge by developing a uniplex SNP genotyping tool called KASP (KBiosciences Competitive Allele-Specific PCR also known as Kompetitive Allele-Specific PCR) (Mir et al., 2013; Neelam, Brown-Guedira, & Huang, 2013; Semagn, Babu, & Hearne, 2014). The KASP genotyping device is a homogeneous fluorescent endpoint genotyping technology that was developed by KBiosciences and later acquired by LGC Genomics in 2011. KASP is a simpler, cheaper, and more compact way to evaluate both SNP and insertion–deletion genotypes among the available uniplex systems (Semagn et al., 2014). GBS or array-based systems may be used to build KASP assays using trait-associated SNP flanking sequences (50-bp upstream and 50-bp downstream around the SNP variant position) (Dereeper, Homa, & Andres, 2015). Here you can find a detailed procedure/protocol for KASP genotyping chemistry, as well as the required tools, software, and reagents, as well as information on designing KASP primers, data production, and data scoring. Allen et al. (2011) used an Avalon/Cadenza doubled haploid mapping population to create a 548 locus genetic linkage map in hexaploid wheat for the first time. SNP genotyping services are available from LGC Genomics directly as well as through the Generation Challenge Program and the Integrated Breeding Platform for a variety of crops, including wheat. The following websites/databases provide information on KASP assays and their mapping to wheat chromosomes:

1. LGC Genomics wheat panel (<http://www.lgcgroup.com/wheat/#.VfMk3q10y70>)
2. CerealsDB KASP SNPs database (<http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php>; https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/kasp_mapped_snps.php)
3. Integrated Breeding Platform (<https://www.integratedbreeding.net/482/communities/genomics-crop-info/crop-information/gcp-kasparnsnp-marker>)
4. LGC's online wheat genotyping ([https://biosearch-cdn.azureedge.net/assetsv6/Wheat-poster-Key-trait screening.pdf](https://biosearch-cdn.azureedge.net/assetsv6/Wheat-poster-Key-trait%20screening.pdf); https://www.researchgate.net/institution/LGC_Biosearch_Technologies2/post/58458fbfd332d599f0c2991_KASPR_Genotyping_Markers_for_Key_Wheat_Traits)

Rasheed et al. (2016) used a panel of 300 cultivars and four RIL populations to validate 70 KASP-based assays of functional markers for agronomic, disease resistance, drought tolerance, preharvest sprouting tolerance, and end-use efficiency traits in wheat. The validated KASP assays related to (1) agronomic traits, including *Ppd-B1*, *Ppd-D1*, *VRN-A1*, *VRN-B1*, *VRN-D1*, *Rht-B1*, *Rht-D1*, *TaCwi-5D*, *TaGS-D1*, *TaTGW6-3A*, *TaGASR-A1*, *TaSus2-2B*, *TaCKX-D1*, and *TaMoc1-7A*; (2) disease resistance, including Lr34TCCIND and Lr34jagger for *Lr34*; (3) drought tolerance, including *TaDreb-B1*, *1-feh w3*, and *TaCwi-4A*; (4) preharvest sprouting tolerance, including *TaPHS1*, *TaSdr-B1*, *TaVp-1B*, and *TaMFT-A1*; and (5) end-use quality comprising *Glu-A1*, *Glu-B1*, *Glu-D1*, *Pina-D1*, *Pinb-D1*, *Pinb-B2*, *Ppo-A1*, *Ppo-D1*, *Psy-A1*, *Psy-B1*, *Psy-D1*, and *Zds-A1* were used for function polymorphism (Rasheed et al., 2016).

Following MAS in wheat genetic improvement, KASP-based SNP markers can be used to pyramid favorable genes/alleles after confirmation. In addition, for genetic research, the KASP marker method has been used in pigeon pea (Saxena et al., 2012), chickpea (Hiremath et al., 2012), Indica rice (Pariasca-Tanaka et al., 2015; Steele et al., 2018), and Japonica rice (Cheon, Baek, Cho, Jeong, & Lee, 2018). The KASP framework allows users to configure a selection of trait-linked SNPs for genotyping and further validation on a panel of wheat germplasm. In addition to the KASP genotyping system, the TaqMan assay (Woodward, 2014), semithermal asymmetric reverse PCR (Long, Chao, Ma, Xu, & Qi, 2016), Amplifluor SNP genotyping system (Jatayev et al., 2017), and RNase H2 enzyme-based amplification (rhAmp) (<https://eu.idtdna.com/pages/products/qpcr-and-pcr/genotyping/rhamp-snp-genotyping>). All five strategies include allele-specific uniplex genotyping platforms with superior chemistry and scalability without sacrificing cost or data throughput (Ayalew et al., 2019; Broccanello et al., 2018; Rasheed et al., 2017).

20.8 Conclusions and future directions

DNA sequencing and genotyping methods have clearly progressed and are now one of the most successful breeding techniques for identifying beneficial alleles that contribute to phenotype variation. Wheat genome sequencing data are being generated at a faster and cheaper pace thanks to the continued help of new technology. The discovery of a larger number of genome-wide SNPs is likely to have the greatest effect on the production of secret variation, particularly near chromosome centromeric regions. The current problems will most likely shift away from wheat genome research and toward the correlation of sequence variance with economically significant traits. High-resolution mapping and cloning of major QTLs will be accelerated by the introduction of ultrahigh-density SNP maps (over 100K markers). Furthermore, combining related QTLs from different experiments and predicting a meta-QTL will further improve the role of QTLs and their associated genes. As a result, user-friendly practical assays based on candidate genes can be used to target alleles in wheat marker-assisted breeding programs. Future study is expected to emerge as molecular marker technology advances, making them a more useful and efficient breeding tool.

References

- Agarwal, M., Shrivastava, N., & Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Reports*, 27, 617–631.
- Ahmad, I., Ali, N., & Ahmad, H. (2017). Association mapping of root traits for drought tolerance in bread wheat. In R. Wanyera, & J. Owuoch (Eds.), *Wheat improvement, management and utilization* (pp. 39–57). London: InTech.
- Ain, Q. U., Rasheed, A., Anwar, A., Mahmood, T., Imtiaz, M., Mahmood, T., Xia, X., et al. (2015). Genome-wide association for grain yield under rainfed conditions in historical wheat cultivars from Pakistan. *Frontiers in Plant Science*, 6, 743.
- Alipour, H., Bai, G., Zhang, G., Bihanta, M. R., Mohammadi, V., & Peyghambari, S. A. (2019). Imputation accuracy of wheat genotyping-by-sequencing (GBS) data using barley and wheat genome references. *PLoS One*, 14(1), e0208614.
- Allen, A. M., Barker, G. L. A., Berry, S. T., Coghill, J. A., Gwilliam, R., Kirby, S., ... McKenzie, N. (2011). Transcript-specific, single nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, 9, 1086–1099.
- Allen, A. M., Barker, G. L., Wilkinson, P., BurrIDGE, A., Winfield, M., Coghill, J., Uauy, C., et al. (2013). Discovery and development of exome-based, codominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, 11, 279–295.
- Allen, A. M., Winfield, M. O., BurrIDGE, A. J., Downie, R. C., Benbow, H. R., Barker, G. L., et al. (2017). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, 15, 390–401.
- Alomari, D. Z., Eggert, K., von Wirén, N., Polley, A., Plieske, J., Ganal, M. W., ... Röder, M. S. (2019). Whole-genome association mapping and genomic prediction for iron concentration in wheat grains. *International Journal of Molecular Sciences*, 20(1), 76.
- Amom, T., & Nongdam, P. (2017). The use of molecular marker methods in plants: A review. *International Journal of Current Research and Review*, 9(17), 1–7.

- Ashikawa, I., Abe, F., & Nakamura, S. (2010). Ectopic expression of wheat and barley *DOG1*-like genes promotes seed dormancy in Arabidopsis. *Plant Science (Shannon, Ireland)*, *179*(5), 536–542.
- Ayalew, H., Tsang, P. W., Chu, C., Wang, J., Liu, S., Chen, C., et al. (2019). Comparison of TaqMan, KASP and rhAmp SNP genotyping platforms in hexaploid wheat. *PLoS One*, *14*(5), e0217222.
- Bajgain, P., Rouse, M. N., Tsilo, T. J., Macharia, G. K., Bhavani, S., et al. (2016). Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. *PLoS One*, *11*, e0155760.
- Beales, J., Turner, A., Griffiths, S., Snape, J. W., & Laurie, D. A. (2007). A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (*Triticum aestivum* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *115*(5), 721–733.
- Belete, T. (2018). Improvement of crop plants by molecular markers. *JOJ Hortic Arboric*, *1*(5), 555572.
- Berkman, P. J., Visendi, P., Lee, H. C., Stiller, J., Manoli, S., Lorenc, M. T., et al. (2013). Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal*, *11*, 564–571.
- Bhatia, D., Wing, R. A., & Singh, K. (2013). Genotyping by sequencing, its implications and benefits. *Crop Improvement*, *40*(2), 101–111.
- Bhatta, M., Morgounov, A., Belamkar, V., & Baenziger, P. S. (2018a). Genome-wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic hexaploid wheat. *International Journal of Molecular Sciences*, *19*, 3011.
- Bhatta, M., Morgounov, A., Belamkar, V., Yorgancılar, A., & Baenziger, P. S. (2018b). Genome-wide association study reveals favorable alleles associated with common bunt resistance in synthetic hexaploid wheat. *Euphytica*, *214*, 200.
- Bohn, M., Utz, H. F., & Melchinger, A. E. (1999). Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. *Crop Science*, *39*, 228–237.
- Brocanello, C., Chiodi, C., Funk, A., McGrath, J. M., Panella, L., & Stevanato, P. (2018). Comparison of three PCR-based assays for SNP genotyping in plants. *Plant Methods*, *14*(1), 28.
- Bryan, G. J., Collins, A. J., Stephenson, P., Orry, A., Smith, J. B., & Gale, M. D. (1997). Isolation and characterization of microsatellites from hexaploid bread wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *94*, 557–563.
- Burkhamer, R. L., Lanning, S. P., Martens, R. J., Martin, J. M., & Talbert, L. E. (1998). Predicting progeny variance from parental divergence in hard red spring wheat. *Crop Science*, *38*(1), 243–248.
- Cabral, A. L., Jordan, M. C., McCartney, C. A., You, F. M., Humphreys, D. G., MacLachlan, R., & Pozniak, C. J. (2014). Identification of candidate genes, regions and markers for pre-harvest sprouting resistance in wheat (*Triticum aestivum* L.). *BMC Plant Biology*, *14*(1), 340.
- Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., . . . Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences of the USA*, *110*(20), 8057–8062.
- Chao, S., Sharp, P. J., Worland, A. J., Warham, E. J., Koebner, R. M. D., & Gale, M. D. (1989). RFLP-based genetic maps of wheat homoeologous group 7 chromosomes. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *78*, 495–504.
- Cheon, K. S., Baek, J., Cho, Y., Jeong, Y.-M., Lee, Y.-Y., et al. (2018). Single nucleotide polymorphism (SNP) discovery and kompetitive allele-specific PCR (KASP) marker development with Korean japonica rice varieties. *Plant Breeding and Biotechnology*, *6*(4), 391–403.
- Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., & Jackson, S. A. (2015). Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. *Molecular Plant*, *8*, 831–846.
- Cloutier, S., McCallum, B. D., Loutre, C., Banks, T. W., Wicker, T., Feuillet, C., . . . Jordan, M. C. (2007). Leaf rust resistance gene *Lr1*, isolated from bread wheat (*Triticum aestivum* L.) is a member of the large *psr567* gene family. *Plant Molecular Biology*, *65*, 93–106.
- Cubizolles, N., Rey, E., Choulet, F., Rimbart, H., Laugier, C., Balfourier, F., . . . Paux, E. (2016). Exploiting the repetitive fraction of the wheat genome for high-throughput single nucleotide polymorphism discovery and genotyping. *Plant Genome*, *9*, 1–11.
- Cui, F., Zhang, N., Fan, X. L., Zhang, W., Zhao, C. H., Yang, L. J., . . . Ji, J. (2017). Utilization of a Wheat660K SNP array-derived high-density genetic map for high-resolution mapping of a major QTL for kernel number. *Scientific Reports*, *7*(1), 3788.
- Davey, J. W., & Blaxter, M. L. (2011). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, *9*, 416–423.
- De Bustos, A., Rubio, P., & Jouve, N. (2001). Characterisation of two gene subunits on the 1R chromosome of rye as orthologs of each of the *Glu-1* genes of hexaploid wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *103*(5), 733–742.
- Dereeper, A., Homa, F., Andres, G., et al. (2015). SNIPlay3: A web-based application for exploration and large scale analyses of genomic variations. *Nucleic Acids Research*, *43*, 295–300.
- Deschamps, S., Llaca, V., & May, G. D. (2012). Genotyping-by-sequencing in plants. *Biology*, *1*(3), 460–483.
- Devos, K. M., & Gale, M. D. (1993). Extended genetic maps of the homoeologous group 3 chromosomes of wheat, rye and barley. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *85*, 649–652.
- Devos, K. M., Moore, G., & Gale, M. D. (1995). Conservation of marker synteny during evolution. *Euphytica*, *85*, 367–372.
- Distelfeld, A., Cakmak, I., Peleg, Z., Ozturk, L., Yazici, A. M., Budak, H., . . . Fahima, T. (2007). Multiple QTL-effects of wheat *Gpc-B1* locus on grain protein and micronutrient concentrations. *Physiologia Plantarum*, *129*(3), 635–643.
- Dong, C. H., Xia, X. C., Zhang, L. P., & He, Z. H. (2012). Allelic variation at the *TaZds-A1* locus on wheat chromosome 2A and development of a functional marker in common wheat. *Journal of Integrative Agriculture*, *11*(7), 1067–1074.
- Dubcovsky, J., & Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science (New York, N.Y.)*, *316*, 1862–1866.
- Edwards, K. J., Reid, A. L., Coghill, J. A., Berry, S. T., & Barker, G. L. (2009). Multiplex single nucleotide polymorphism (SNP)-based genotyping in allohexaploid wheat using padlock probes. *Plant Biotechnology Journal*, *7*, 375–390.

- Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., & Akhunov, E. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Science (Shannon, Ireland)*, *270*, 123–130.
- Ellis, M., Spielmeier, W., Gale, K., Rebetzke, G., & Richards, R. (2002). Perfect markers for the *Rht-B1b* and *Rht-D1b* dwarfing genes in wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *105*, 1038–1042.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, *6*, e19379.
- FAO. (2017). FAOSTAT. <http://www.fao.org/faostat/en/>.
- Faris, J. D., Zhang, Z., Lu, H., Lu, S., Reddy, L., Cloutier, S., . . . Oliver, R. P. (2010). A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. *Proceedings of the National Academy of Sciences of the USA*, *107*(30), 13544–13549.
- Feuillet, C., Travella, S., Stein, N., Albar, L., Nublát, A., & Keller, B. (2003). Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proceedings of the National Academy of Sciences of the USA*, *100*(25), 15253–15258.
- Forrest, K., Pujol, V., Bulli, P., Pumphery, M., Wellings, C., Herrera-Foessel, S., . . . Spielmeier, W. (2014). Development of a SNP marker assay for the *Lr67* gene of wheat using a genotyping by sequencing approach. *Molecular Breeding*, *34*, 2109–2118.
- Fu, D., Szucs, P., Yan, L., Helguera, M., Skinner, J. S., Von Zitzewitz, J., . . . Dubcovsky, J. (2005). Large deletions within the first intron in *VRN-1* are associated with spring growth habit in barley and wheat. *Molecular Genetics and Genomics: MGG*, *273*(1), 54–65.
- Fu, D., Uauy, C., Distelfeld, A., Blechl, A., Epstein, L., Chen, X., . . . Dubcovsky, J. (2009). A kinase-START gene confers temperature-dependent resistance to wheat stripe rust. *Science (New York, N.Y.)*, *323*, 1357–1360.
- Gale, K. R. (2005). Diagnostic DNA markers for quality traits in wheat. *Journal of Cereal Science*, *41*, 181–192.
- Ganal, M. W., Altmann, T., & Roder, M. S. (2009). SNP identification in crop plants. *Current Opinion in Plant Biology*, *12*, 211–217.
- Gao, F., Liu, J., Yang, L., Wu, X., Xiao, Y., Xia, X., et al. (2016a). Genome-wide linkage mapping of QTL for physiological traits in a Chinese wheat population using the 90K SNP array. *Euphytica*, *209*, 789–804.
- Gao, L., Kielsmeier-Cook, J., Bajgain, P., Zhang, X., Chao, S., Rouse, M. N., & Anderson, J. A. (2015). Development of genotyping by sequencing (GBS)- and array-derived SNP markers for stem rust resistance gene *Sr42*. *Molecular Breeding*, *35*, 207.
- Gao, L., Turner, M. K., Chao, S., Kolmer, J., & Anderson, J. A. (2016b). Genome-wide association study of seedling and adult plant leaf rust resistance in elite spring wheat breeding lines. *PLoS One*, *11*, e0148671.
- Gao, L., Zhao, G., Huang, D., & Jia, J. (2017). Candidate loci involved in domestication and improvement detected by a published 90K wheat SNP array. *Scientific Reports*, *7*, 44530.
- Garcia, M., Eckermann, P., Haefele, S., Satija, S., Sznajder, B., Timmins, A., et al. (2019). Genome-wide association mapping of grain yield in a diverse collection of spring wheat (*Triticum aestivum* L.) evaluated in southern Australia. *PLoS One*, *14*(2), e0211730.
- Garg, B., Lata, C., & Prasad, M. (2012). A study of the role of gene *TaMYB2* and an associated SNP in dehydration tolerance in common wheat. *Molecular Biology Reports*, *39*(12), 10865–10871.
- Gautier, M. F., Aleman, M. E., Guirao, A., Marion, D., & Joudrier, P. (1994). *Triticum aestivum* puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. *Plant Molecular Biology*, *25*(1), 43–57.
- Geng, H., Shi, J., Fuerst, E. P., Wei, J., & Morris, C. F. (2019). Physical mapping of peroxidase genes and development of functional markers for *TaPod-D1* on bread wheat chromosome 7D. *Frontiers in Plant Science*, *10*, 523.
- Geng, H., Xia, X., Zhang, L., Qu, Y., & He, Z. (2012). Development of functional markers for a lipoxygenase gene *TaLox-B1* on chromosome 4BS in common wheat. *Crop Science*, *52*(2), 568–576.
- Goyal, A., Bandopadhyay, R., Sourdille, P., Endo, T. R., Balyan, H. S., & Gupta, P. K. (2005). Physical molecular maps of wheat chromosomes. *Functional & Integrative Genomics*, *5*, 260–263.
- Gunderson, K. L. (2009). Whole-genome genotyping on bead arrays. *Methods in Molecular Biology*, *529*, 197–213.
- Gupta, P. K., Balyan, H. S., Edwards, K. J., Isaac, P., Korzun, V., Roder, M. S., . . . Leroy, P. (2002). Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *105*, 413–422.
- Gupta, P. K., Balyan, H. S., & Gahlaut, V. (2017). QTL analysis for drought tolerance in wheat: Present status and future possibilities. *Agronomy*, *7*, 5.
- Gupta, P. K., Mir, R. R., Mohan, A., & Kumar, J. (2008a). Wheat genomics: Present status and future prospects. *International Journal of Plant Genomics*, *2008*, 896451.
- Gupta, P. K., Rustgi, S., & Mir, R. R. (2008b). Array-based high-throughput DNA markers for crop improvement. *Heredity*, *101*, 5–18.
- Gupta, P. K., Rustgi, S., & Mir, R. R. (2013). Array-based high-throughput DNA markers and genotyping platforms for cereal genetics and genomics. In P. K. Gupta, & R. K. Varshney (Eds.), *Cereal genomics II* (pp. 11–55). Dordrecht: Springer.
- Gupta, P. K., & Varshney, R. K. (2000). The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica*, *113*, 163–185.
- Gupta, P. K., Varshney, R. K., Sharma, P. C., & Ramesh, B. (1999). Molecular markers and their applications in wheat breeding. *Plant Breeding*, *118* (5), 369–390.
- Guzmán, C., Caballero, L., Martín, M. A., & Alvarez, J. B. (2012). Molecular characterization and diversity of the *Pina* and *Pinb* genes in cultivated and wild diploid wheat. *Molecular Breeding*, *30*(1), 69–78.
- Hanif, M., Gao, F., Liu, J., Wen, W., Zhang, Y., Rasheed, A., . . . Cao, S. (2016). *TaTGW6-A1*, an ortholog of rice *TGW6*, is associated with grain weight and yield in bread wheat. *Molecular Breeding*, *36*(1), 1.
- He, C., Holme, J., & Anthony, J. (2014). SNP genotyping: The KASP assay. In D. Fleury, & R. Whitford (Eds.), *Crop breeding. Methods in molecular biology (methods and protocols)* (Vol. 1145, pp. 75–86). New York: Humana Press.

- He, X. Y., He, Z. H., Ma, W., Appels, R., & Xia, X. C. (2009). Allelic variants of *phytoene synthase 1 (Psy1)* genes in Chinese and CIMMYT wheat cultivars and development of functional markers for flour colour. *Molecular Breeding*, 23(4), 553–563.
- He, X. Y., He, Z. H., Zhang, L. P., Sun, D. J., Morris, C. F., Fuerst, E. P., & Xia, X. C. (2007). Allelic variation of *polyphenol oxidase (PPO)* genes located on chromosomes 2A and 2D and development of functional markers for the *PPO* genes in common wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 115(1), 47–58.
- Himi, E., & Noda, K. (2005). Red grain colour gene (R) of wheat is a Myb-type transcription factor. *Euphytica*, 143(3), 239–242.
- Hiremath, P. J., Kumar, A., Penmetsa, R. V., Farmer, A., Schlueter, J. A., Chamathi, S. K., et al. (2012). Large-scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. *Plant Biotechnology Journal*, 10, 716–732.
- Hu, M. J., Zhang, H. P., Liu, K., Cao, J. J., Wang, S. X., Jiang, H., . . . Sun, G. L. (2016). Cloning and characterization of *TaTGW-7A* gene associated with grain weight in wheat via SLAF-seq-BSA. *Frontiers in Plant Science*, 7, 1902.
- Huang, L., Brooks, S. A., Li, W., Fellers, J. P., Trick, H. N., & Gill, B. S. (2003). Map-based cloning of leaf rust resistance gene *Lr21* from the large and polyploid genome of bread wheat. *Genetics*, 164(2), 655–664.
- Huang, Y. F., Poland, J. A., Wight, C. P., Jackson, E. W., & Tinker, N. A. (2014). Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *PLoS One*, 9, e102448.
- Hurni, S., Brunner, S., Buchmann, G., Herren, G., Jordan, T., Krukowski, P., . . . Keller, B. (2013). Rye *Pm8* and wheat *Pm3* are orthologous genes and show evolutionary conservation of resistance function against powdery mildew. *The Plant Journal: For Cell and Molecular Biology*, 76(6), 957–969.
- Hussain, S. S., & Qamar, R. (2007). Wheat genomics: Challenges and alternative strategies. *Proceedings of the Pakistan Academy of Sciences*, 44, 305–327.
- Jamil, M., Ali, A., Gul, A., Ghafoor, A., Napar, A. A., Ibrahim, A. M., . . . Mujeeb-Kazi, A. (2019). Genome-wide association studies of seven agronomic traits under two sowing conditions in bread wheat. *BMC Plant Biology*, 19(1), 149.
- Jatayev, S., Kurishbayev, A., Zotova, L., Khasanova, G., Serikbay, D., Zhubatkanov, A., . . . Langridge, P. (2017). Advantages of Amplifluor-like SNP markers over KASP in plant genotyping. *BMC Plant Biology*, 17(2), 254.
- Jefferies, S. P., Pallotta, M. A., Paull, J. G., Karakousis, A., Kretschmer, J. M., Manning, S., . . . Chalmers, K. J. (2000). Mapping and validation of chromosome regions conferring boron toxicity tolerance in wheat (*Triticum aestivum*). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 101, 767–777.
- Jiang, G. L. (2013). Molecular markers and marker-assisted breeding in plants, plant breeding from laboratories to fields. In S. B. Andersen (Ed.), *Plant breeding from laboratories to fields* (p. 52583). Rijeka: InTech. Publishers.
- Jiang, Q., Hou, J., Hao, C., Wang, L., Ge, H., Dong, Y., & Zhang, X. (2011). The wheat (*Triticum aestivum*) *sucrose synthase 2* gene (*TaSus2*) active in endosperm development is associated with yield traits. *Functional & Integrative Genomics*, 11(1), 49–61.
- Jordan, K., Wang, S., Lun, Y., Gardiner, L.-J., MacLachlan, R., Hucl, P., et al. (2015). A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biology*, 16, 48.
- Jordan, T., Seeholzer, S., Schwizer, S., Toller, A., Somssich, I. E., & Keller, B. (2011). The wheat *Mla* homologue *TmMla1* exhibits an evolutionarily conserved function against powdery mildew in both wheat and barley. *The Plant Journal: For Cell and Molecular Biology*, 65(4), 610–621.
- Kesawat, M. S., & Das, B. K. (2009). Molecular markers: It's application in crop improvement. *Journal of Crop Science and Biotechnology*, 12(4), 168–178.
- Klymiuk, V., Yaniv, E., Huang, L., Raats, D., Fatiukha, A., Chen, S., . . . Chang, W. (2018). Cloning of the wheat *Yr15* resistance gene sheds light on the plant tandem kinase-pseudokinase family. *Nature Communications*, 9(1), 3735.
- Korzun, V., & Ebmeyer, E. (2003). Molecular markers and their applications in wheat breeding. In N. E. Pogna, M. Romano, E. A. Pogna, & G. Galterio (Eds.), *Proceedings 10th international wheat genetics symposium, Istituto Sperimentale per la Cerealicoltura* (Vol. 1, pp. 140–143). Rome, Italy.
- Korzun, V., Roder, M. S., Worland, A. J., & Borner, A. (1997). Application of microsatellite markers to distinguish inter-varietal chromosome substitution lines of wheat (*Triticum aestivum* L.). *Euphytica*, 95, 149–155.
- Krattinger, S. G., Lagudah, E. S., Spielmeier, W., Singh, R. P., Huerta-Espino, J., McFadden, H., . . . Keller, B. (2009). A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science (New York, N.Y.)*, 323, 1360–1363.
- Kumar, S., Goyal, A., Mohan, A., Balyan, H. S., & Gupta, P. K. (2013). An integrated physical map of simple sequence repeats in bread wheat. *Australian Journal of Crop Science*, 7(4), 460–468.
- Kumar, S., Knox, R. E., Singh, A. K., DePauw, R. M., Campbell, H. L., Isidro-Sanchez, J., . . . Fedak, G. (2018). High-density genetic mapping of a major QTL for resistance to multiple races of loose smut in a tetraploid wheat cross. *PLoS One*, 13(2), e0192261.
- Kumar, S., Phogat, B. S., Vikas, V. K., Sharma, A. K., Saharan, M. S., Singh, A. K., et al. (2019). Mining of Indian wheat germplasm collection for adult plant resistance to leaf rust. *PLoS One*, 14(3), e0213468.
- Lagudah, E. S., Krattinger, S. G., Herrera-Foessel, S., et al. (2009). Gene-specific markers for the wheat gene *Lr34/Yr18/Pm38*, which confers resistance to multiple fungal pathogens. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 119(5), 889–898.
- Lai, K., Duran, C., Berkman, P. J., Lorenc, M. T., Stiller, J., Manoli, S., . . . Edwards, D. (2012). Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnology Journal*, 10, 743–749.
- Landjeva, S., Korzun, V., & Borner, A. (2007). Molecular markers: Actual and potential contributions to wheat genome characterization and breeding. *Euphytica*, 156, 271–296.

- Langridge, P., & Chalmers, K. (2004). Biotechnology in agriculture and forestry. In H. Lorz, & G. Wenzel (Eds.), *Molecular marker systems* (Vol. 55). Berlin Heidelberg: Springer-Verlag.
- Langridge, P., Lagudah, E. S., Holton, T. A., Appels, A., Sharp, P., & Chalmers, K. J. (2001). Trends in genetic and genome analyses in wheat: A review. *Australian Journal of Agricultural Research*, 52, 1043–1077.
- Li, B., Li, Q., Mao, X., Li, A., Wang, J., Chang, X., ... Jing, R. (2016). Two novel *AP2/EREBP* transcription factor genes *TaPARG* have pleiotropic functions on plant architecture and yield-related traits in common wheat. *Frontiers in Plant Science*, 7, 1191.
- Li, F., Wen, W., Liu, J., et al. (2019). Genetic architecture of grain yield in bread wheat based on genome-wide association studies. *BMC Plant Biology*, 19, 168.
- Li, G., Wang, Y., Chen, M. S., Edeae, E., Poland, J., Akhunov, E., ... Yan, L. (2015b). Precisely mapping a major gene conferring resistance to hessian fly in bread wheat using genotyping-by-sequencing. *BMC Genomics*, 16(1), 108.
- Li, H., Vikram, P., Singh, R. P., Kilian, A., Carling, J., Song, J., ... Singh, S. (2015a). A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics*, 16(1), 216.
- Lin, M., Cai, S., Wang, S., Liu, S., Zhang, G., & Bai, G. (2015). Genotyping-by-sequencing (GBS) identified SNP tightly linked to QTL for pre-harvest sprouting resistance. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 128(7), 1385–1395.
- Lindsay, M. P., Lagudah, E. S., Hare, R. A., & Munns, R. (2004). A locus for sodium exclusion (*Nax1*), a trait for salt tolerance, mapped in durum wheat. *Functional Plant Biology: FPB*, 31(11), 1105–1114.
- Liu, J., He, Z., Rasheed, A., et al. (2017). Genome-wide association mapping of black point reaction in common wheat (*Triticum aestivum* L.). *BMC Plant Biology*, 17, 220.
- Liu, J., Xu, Z., Fan, X., Zhou, Q., Cao, J., Wang, F., ... Wang, T. (2018). A genome-wide association study of wheat spike related traits in China. *Frontiers in Plant Science*, 9, 1584.
- Liu, S., Assanga, S. O., Dhakal, S., Gu, X., Tan, C. T., Yang, Y., et al. (2016). Validation of chromosomal locations of 90K array single nucleotide polymorphisms in US wheat. *Crop Science*, 56(1), 364–373.
- Liu, S., Chao, S., & Anderson, J. A. (2008). New DNA markers for high molecular weight glutenin subunits in wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 118(1), 177.
- Liu, W., Frick, M., Huel, R., Nykiforuk, C. L., Wang, X., Gaudet, D. A., ... Kang, Z. (2014). The stripe rust resistance gene *Yr10* encodes an evolutionary-conserved and unique CC–NBS–LRR sequence in wheat. *Molecular Plant*, 7(12), 1740–1755.
- Long, Y. M., Chao, W. S., Ma, G. J., Xu, S. S., & Qi, L. L. (2016). An innovative SNP genotyping method adapting to multiple platforms and throughputs. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 130, 597–607.
- Lorz, H., & Wenzel, G. (2004). *Molecular marker systems in plant breeding and crop improvement* (p. 476) Germany: Springer Verlag.
- Ma, D., Yan, J., He, Z., Wu, L., & Xia, X. (2012). Characterization of a cell wall invertase gene *TaCwi-A1* on common wheat chromosome 2A and development of functional markers. *Molecular Breeding*, 29(1), 43–52.
- Ma, L., Li, T., Hao, C., Wang, Y., Chen, X., & Zhang, X. (2016). *TaGS5-3A*, a grain size gene selected during wheat improvement for larger kernel and yield. *Plant Biotechnology Journal*, 14(5), 1269–1280.
- Ma, Z. Q., Roder, M., & Sorrells, M. E. (1996). Frequencies and sequence characteristics of di-, tri- and tetra- nucleotide microsatellites in wheat. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 39, 123–130.
- Mago, R., Zhang, P., Vautrin, S., Simkova, H., Bansal, U., Luo, M. C., ... Jin, Y. (2015). The wheat *Sr50* gene reveals rich diversity at a cereal disease resistance locus. *Nature Plants*, 1(12), 15186.
- Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, 387–402.
- Massa, A. N., Morris, C. F., & Gill, B. S. (2004). Sequence diversity of puroindoline-a, puroindoline-b, and the grain softness protein genes in *Aegilops tauschii* Coss. *Crop Science*, 44(5), 1808–1816.
- Mir, R. R., Hiremath, P. J., Riera-Lizarazu, O., & Varshney, R. K. (2013). Evolving molecular marker technologies in plants: From RFLPs to GBS. In T. Lubberstedt, & R. K. Varshney (Eds.), *Diagnostics in plant breeding* (pp. 229–247). New York: Springer.
- Mir, R. R., & Varshney, R. K. (2013). Future prospects of molecular markers in plants. In R. J. Henry (Ed.), *Molecular markers in plants* (pp. 169–190). Oxford: Blackwell Publishing Ltd.
- Moore, J. W., Herrera-Foessel, S., Lan, C., Schnippenkoetter, W., Ayliffe, M., Huerta-Espino, J., ... Kong, X. (2015). A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. *Nature Genetics*, 47(12), 1494.
- Moose, S. P., & Mumm, R. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology*, 147, 969–977.
- Mujeeb-Kazi, A., Kazi, A. G., Dundas, I., Rasheed, A., Ogbonnaya, F., Kishii, M., et al. (2013). Genetic diversity for wheat improvement as a conduit to food security. *Advances in Agronomy*, 122, 179–257.
- Murai, J., Taira, T., & Ohta, D. (1999). Isolation and characterization of the three waxy genes encoding the granule-bound starch synthase in hexaploid wheat. *Gene*, 234(1), 71–79.
- Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Dogan, Y., Comertpay, G., Yildiz, M., ... Baloch, F. S. (2018). DNA molecular markers in plant breeding: Current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment*, 32(2), 261–285.
- Nakamura, S., Abe, F., Kawahigashi, H., Nakazono, K., Tagiri, A., Matsumoto, T., ... Mori, M. (2011). A wheat homolog of Mother of FT and TFL1 acts in the regulation of germination. *The Plant Cell*, 23(9), 3215–3229.
- Neelam, K., Brown-Guedira, G., & Huang, L. (2013). Development and validation of a breeder-friendly KASPar marker for wheat leaf rust resistance locus *Lr21*. *Molecular Breeding*, 31(1), 233–237.

- Negro, S. S., Millet, E. J., Madur, D., et al. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biology*, *19*, 318.
- Ogbonnaya, F. C., Subrahmanyam, N. C., Moullet, O., De Majnik, J., Eagles, H. A., Brown, J. S., . . . Lagudah, E. S. (2001). Diagnostic DNA markers for cereal cyst nematode resistance in bread wheat. *Australian Journal of Agricultural Research*, *52*(12), 1367–1374.
- Ohnishi, N., Himi, E., Yamasaki, Y., & Noda, K. (2008). Differential expression of three ABA-insensitive five binding protein (AFP)-like genes in wheat. *Genes & Genetic Systems*, *83*(2), 167–177.
- Pan, Z., Song, W., Meng, F., Xu, L., Liu, B., & Zhu, J. (2004). Characterization of genes encoding wheat grain hardness from Chinese cultivar GaoCheng 8901. *Cereal Chemistry*, *81*(2), 287–289.
- Pandey, B., Sharma, P., Pandey, D. M., Sharma, I., & Chatrath, R. (2013). Identification of new aquaporin genes and single nucleotide polymorphism in bread wheat. *Evolutionary Bioinformatics Online*, *9*, 437–452.
- Pariasca-Tanaka, J., Lorieux, M., He, C., McCouch, S., Thomson, M. J., & Wissuwa, M. (2015). Development of a SNP genotyping panel for detecting polymorphisms in *Oryza glaberrima* × *O. sativa* interspecific crosses. *Euphytica*, *201*, 67–78.
- Parker, G. D., Chalmers, K. J., Rathjen, A. J., & Langridge, P. (1999). Mapping loci associated with milling yield in wheat (*Triticum aestivum* L.). *Molecular Breeding*, *5*, 561–568.
- Paux, E., Sourdille, P., Mackay, I., & Feuillet, C. (2011). Sequence-based marker development in wheat: Advances and applications to breeding. *Biotechnology Advances*, *30*, 1071–1088.
- Pariyannan, S., Bansal, U., Bariana, H., Deal, K., Luo, M. C., Dvorak, J., & Lagudah, E. (2014). Identification of a robust molecular marker for the detection of the stem rust resistance gene *Sr45* in common wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *127*(4), 947–955.
- Pariyannan, S., Moore, J., Ayliffe, M., Bansal, U., Wang, X., Huang, L., . . . Mago, R. (2013). The gene *Sr33*, an ortholog of barley *Mla* genes, encodes resistance to wheat stem rust race Ug99. *Science (New York, N.Y.)*, *341*, 786–788.
- Picoult-Newberg, L., Ideker, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A., & Boyce-Jacino, M. (1999). Mining SNPs from EST databases. *Genome Research*, *9*(167), 174.
- Plaschke, J., Ganal, M., & Roder, M. S. (1995). Detection of genetic diversity in closely related bread wheat using microsatellite markers. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *91*, 1001–1007.
- Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012a). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, *7*, e32253.
- Poland, J. A., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012b). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome*, *5*, 103–113.
- Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, *5*, 92–102.
- Prasad, M., Varshney, R. K., Roy, J. K., Balyan, H. S., & Gupta, P. K. (2000). The use of microsatellites for detecting DNA polymorphism, genotype identification and genetic diversity in wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *100*, 584–592.
- Przewieslik-Allen, A. M., Burridge, A. J., Wilkinson, P. A., Winfield, M. O., Shaw, D. S., McAusland, L., . . . Barker, G. L. A. (2019). Developing a high-throughput SNP-based marker system to facilitate the introgression of traits from *Aegilops* species into bread wheat (*Triticum aestivum*). *Frontiers in Plant Science*, *9*, 1993.
- Randhawa, H. S., Popovic, Z., Menzies, J., Knox, R., & Fox, S. (2009). Genetics and identification of molecular markers linked to resistance to loose smut (*Ustilago tritici*) race T33 in durum wheat. *Euphytica*, *169*(2), 151–157.
- Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., & He, Z. (2017). Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives. *Molecular Plant*, *10*, 1047–1064.
- Rasheed, A., Mujeeb-Kazi, A., Ogbonnaya, F. C., He, Z. H., & Rajaram, S. (2018a). Wheat genetic resources in the post-genomics era: Promise and challenges. *Annals of Botany*, *121*, 603–616.
- Rasheed, A., Wen, W., Gao, F., Zhai, S., Jin, H., Liu, J., & He, Z. (2016). Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *129*(10), 1843–1860.
- Rasheed, A., & Xia, X. (2019). From markers to genome-based breeding in wheat. *Theoretical and Applied Genetics*, *132*, 767–784. Available from <https://doi.org/10.1007/s00122-019-03286-4>.
- Rawat, N., Pumphrey, M. O., Liu, S., Zhang, X., Tiwari, V. K., Ando, K., . . . Gill, B. S. (2016). Wheat *Fhb1* encodes a chimeric lectin with agglutinin domains and a pore-forming toxin-like domain conferring resistance to Fusarium head blight. *Nature Genetics*, *48*(12), 1576.
- Rehman, S. U., Wang, J., Chang, X., Zhang, X., Mao, X., & Jing, R. (2019). A wheat protein kinase gene TaSnRK2.9–5A associated with yield contributing traits. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *132*(4), 907–919.
- Ren, T., Hu, Y., Tang, Y., Li, C., Yan, B., Ren, Z., . . . Li, Z. (2018). Utilization of a Wheat55K SNP array for mapping of major QTL for temporal expression of the tiller number. *Frontiers in Plant Science*, *9*, 333.
- Rimbert, H., Darrier, B., Navarro, J., Kitt, J., Choulet, F., Leveugle, M., et al. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS One*, *13*(1), e0186329.
- Roder, M. S., Huang, X.-Q., & Ganal, M. W. (2004). Wheat microsatellites: Potential and implications. In H. Lörz, & G. Wenzel (Eds.), *Molecular marker systems in plant breeding and crop improvement. Biotechnology in agriculture and forestry* (Vol. 55, pp. 255–266). Berlin Heidelberg: Springer.
- Roder, M. S., Plaschke, J., Konig, S. U., Borner, A., & Sorrells, M. E. (1995). Abundance, variability and chromosomal location of microsatellite in wheat. *Molecular & General Genetics: MGG*, *246*, 327–333.

- Rustgi, S., Bandopadhyay, R., Balyan, H. S., & Gupta, P. K. (2009). EST-SNPs in bread wheat: Discovery, validation, genotyping and haplotype structure. *Czech Journal of Genetics and Plant Breeding*, *45*, 106–116.
- Saintenac, C., Jiang, D., Wang, S., & Akhunov, E. (2013a). Sequence-based mapping of the polyploid wheat genome. *G3 (Bethesda)*, *3*, 1105–1114.
- Saintenac, C., Lee, W. S., Cambon, F., Rudd, J. J., King, R. C., Marande, W., . . . Hammond-Kosack, K. E. (2018). Wheat receptor-kinase-like protein Stb6 controls gene-for-gene resistance to fungal pathogen *Zymoseptoria tritici*. *Nature Genetics*, *50*(3), 368.
- Saintenac, C., Zhang, W., Salcedo, A., Rouse, M. N., Trick, H. N., Akhunov, E., & Dubcovsky, J. (2013b). Identification of wheat gene *Sr35* that confers resistance to Ug99 stem rust race group. *Science (New York, N.Y.)*, *341*, 783–786.
- Saxena, R. K., Penmetsa, R. V., Upadhyaya, H. D., Kumar, A., Carrasquilla-Garcia, N., Schlueter, J. A., et al. (2012). Large-scale development of cost-effective single-nucleotide polymorphism marker assays for genetic mapping in pigeonpea and comparative mapping in legumes. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *19*, 449–461.
- Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*, 149–161.
- Sehgal, D., Vikram, P., Sansaloni, C. P., Ortiz, C., Pierre, C. S., Payne, T., et al. (2015). Exploring and mobilizing the gene bank biodiversity for wheat improvement. *PLoS One*, *10*(7), e0132112.
- Semagn, K., Babu, R., Hearne, S., et al. (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement. *Molecular Breeding*, *33*, 1–14.
- Shi, G., Zhang, Z., Friesen, T. L., Raats, D., Fahima, T., Bruggeman, R. S., . . . Frenkel, Z. (2016). The hijacking of a receptor kinase-driven pathway by a wheat fungal pathogen leads to disease. *Science Advances*, *2*(10), e1600822.
- Shorinola, O., Bird, N., Simmonds, J., Berry, S., Henriksson, T., Jack, P., . . . Valárik, M. (2016). The wheat *Phs-A1* pre-harvest sprouting resistance locus delays the rate of seed dormancy loss and maps 0.3 cM distal to the *PM19* genes in UK germplasm. *Journal of Experimental Botany*, *67* (14), 4169–4178.
- Sieber, A. N., Longin, C. F. H., Leiser, W. L., & Würschum, T. (2016). Copy number variation of *CBF-A14* at the *Fr-A2* locus determines frost tolerance in winter durum wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *129*(6), 1087–1097.
- Sourdille, P., Singh, S., Cadalen, T., Gina, L., Brown-Guedira, G. G., Qi, L., . . . Bernard, M. (2004). Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (*Triticum aestivum* L.). *Functional & Integrative Genomics*, *4*, 12–25.
- Steele, K. A., Quinton-Tulloch, M. J., Amgai, R. B., Dhakal, R., Khatiwada, S. P., Vyas, D., et al. (2018). Accelerating public sector rice breeding with high-density KASP markers derived from whole genome sequencing of Indica rice. *Molecular Breeding*, *38*, 38.
- Su, Z., Hao, C., Wang, L., Dong, Y., & Zhang, X. (2011). Identification and development of a functional marker of *TaGW2* associated with grain weight in bread wheat (*Triticum aestivum* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *122*(1), 211–223.
- Sun, D. J., He, Z. H., Xia, X. C., Zhang, L. P., Morris, C. F., Appels, R., . . . Wang, H. (2005). A novel STS marker for polyphenol oxidase activity in bread wheat. *Molecular Breeding*, *16*(3), 209–218.
- Tabbitta, F., Pearce, S., & Barneix, A. J. (2017). Breeding for increased grain protein and micronutrient content in wheat: Ten years of the GPC-B1 gene. *Journal of Cereal Science*, *73*, 183–191.
- Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding and Biotechnology*, *2*, 195–212.
- Tommasini, L., Yahiaoui, N., Srichumpa, P., & Keller, B. (2006). Development of functional markers specific for seven *Pm3* resistance alleles and their validation in the bread wheat gene pool. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *114*(1), 165–175.
- Torkamaneh, D., & Belzile, F. (2015). Scanning and filling: Ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole genome resequencing data. *PLoS One*, *10*(7), e0131533.
- Turuspekov, Y., Plieske, J., Ganai, M., Akhunov, E., & Abugalieva, S. (2015). Phylogenetic analysis of wheat cultivars in Kazakhstan based on the wheat 90 K single nucleotide polymorphism array. *Plant Genetic Resources*, *15*, 29–35.
- Uauy, C., Distelfeld, A., Fahima, T., Blechl, A., & Dubcovsky, J. (2006). A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science (New York, N.Y.)*, *314*, 1298–1301.
- Utsugi, S., Nakamura, S., Noda, K., & Maekawa, M. (2008). Structural and functional properties of *Viviparous1* genes in dormant wheat. *Genes & Genetic Systems*, *83*(2), 153–166.
- Varshney, R., Graner, A., & Sorrells, M. E. (2005). Genic microsatellite markers in plants: Features and applications. *Trends in Biotechnology*, *23*, 48–55.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., . . . Zabeau, M. (1995). AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research*, *23*, 4407–4414.
- Wang, J., Wen, W., Hanif, M., Xia, X., Wang, H., Liu, S., . . . He, Z. (2016). *TaELF3-IDL*, a homolog of *ELF3*, is associated with heading date in bread wheat. *Molecular Breeding*, *36*(12), 161.
- Wang, L., Li, G., Peña, R. J., Xia, X., & He, Z. (2010). Development of STS markers and establishment of multiplex PCR for *Glu-A3* alleles in common wheat (*Triticum aestivum* L.). *Journal of Cereal Science*, *51*(3), 305–312.
- Wang, S., Wong, D., Forrest, K., Allen, A., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90000 single nucleotide polymorphism array. *Plant Biotechnology Journal*, *12*, 787–796.
- Wei, B., Jing, R., Wang, C., Chen, J., Mao, X., Chang, X., & Jia, J. (2009). *DreB1* genes in wheat (*Triticum aestivum* L.): Development of functional markers and gene mapping based on SNPs. *Molecular Breeding*, *23*(1), 13–22.
- Wei, J., Geng, H., Zhang, Y., Liu, J., Wen, W., Zhang, Y., . . . He, Z. (2015). Mapping quantitative trait loci for peroxidase activity and developing gene-specific markers for *TaPod-A1* on wheat chromosome 3AL. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *128*(10), 2067–2076.

- Williams, K., Taylor, S., Bogacki, P., Pallotta, M., Bariana, H., & Wallwork, H. (2002). Mapping of the root lesion nematode (*Pratylenchus neglectus*) resistance gene *Rlm1* in wheat. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 104(5), 874–879.
- Winfield, M. O., Allen, A. M., Burrridge, A. J., Barker, G. L., Benbow, H. R., Wilkinson, P. A., et al. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnology Journal*, 14, 1195–1206.
- Woodward, J. (2014). Bi-allelic SNP genotyping using the TaqMan® assay. *Methods in Molecular Biology*, 1145, 67–74.
- Worland, A. J., Borner, A., Korzun, V., Li, W. M., Petrovic, S., & Sayers, E. J. (1998). The influence of photoperiod genes on the adaptability of Euro-pean winter wheats (Reprinted from *Wheat: Prospects for global improvement*, 1998). *Euphytica*, 100, 385–394.
- Yahiaoui, N., Srichumpa, P., Dudler, R., & Keller, B. (2004). Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene *Pm3b* from hexaploid wheat. *The Plant Journal: For Cell and Molecular Biology*, 37(4), 528–538.
- Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T., & Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene *VRN1*. *Proceedings of the National Academy of Sciences of the USA*, 100(10), 6263–6268.
- Zane, L., Bargelloni, L., & Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*, 11, 1–16.
- Zhang, B., Liu, X., Xu, W., Chang, J., Li, A., Mao, X., . . . Jing, R. (2015). Novel function of a putative *MOC1* ortholog associated with spikelet number per spike in common wheat. *Scientific Reports*, 5, 12211.
- Zhang, C., Dong, C., He, X., Zhang, L., Xia, X., & He, Z. (2011). Allelic variants at the *TaZds-D1* locus on wheat chromosome 2DL and their association with yellow pigment content. *Crop Science*, 51(4), 1580–1590.
- Zhang, P., He, Z., Tian, X., Gao, F., Xu, D., Liu, J., . . . Xia, X. (2017b). Cloning of *TaTPP-6AL1* associated with grain weight in bread wheat and development of functional marker. *Molecular Breeding*, 37(6), 78.
- Zhang, W., Chen, S., Abate, Z., Nirmala, J., Rouse, M. N., & Dubcovsky, J. (2017a). Identification and characterization of *Sr13*, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. *Proceedings of the National Academy of Sciences of the USA*, 114(45), 9483–9492.
- Zhang, Y., Miao, X., Xia, X., & He, Z. (2014). Cloning of seed dormancy genes (*TaSdr*) associated with tolerance to pre-harvest sprouting in common wheat and development of a functional marker. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 127(4), 855–866.
- Zhao, J., Abdelsalam, N. R., LKhalaf, L., Chuang, W.-P., Zhao, L., Smith, M. C., . . . Bai, G. (2019). Development of single nucleotide polymorphism markers for the wheat curl mite resistance gene *Cmc4*. *Crop Science*, 59, 1567–1575.
- Zou, J., Semagn, K., Iqbal, M., N'diaye, A., Chen, H., Asif, M., et al. (2016). Mapping QTLs controlling agronomic traits in the 'Attila' × 'CDC go' spring wheat population under organic management using 90K SNP array. *Crop Science*, 57, 365–377.

Omics approaches for biotic, abiotic, and quality traits improvement in potato (*Solanum tuberosum* L.)

Jagesh Kumar Tiwari¹, Tanuja Buckseth¹, Clarissa Challam², Nandakumar Natarajan², Rajesh K. Singh¹ and Manoj Kumar¹

¹ICAR-Central Potato Research Institute, Shimla, Himachal Pradesh, India, ²ICAR-Central Potato Research Institute, Regional Station, Shillong, Meghalaya, India

21.1 Introduction

Potato (*Solanum tuberosum* L.) is one of the world's most important food crops, ranking third in terms of global production just behind rice and wheat (Birch et al., 2012). Limited genetic variation makes the crop vulnerable to disease and insect epidemics. Genetic vulnerability could have devastating effects on mankind, as occurred in Ireland in 1845, when the late blight disease caused by *Phytophthora infestans* destroyed the potato crop. The famine that occurred resulted in death of million Irish people due to starvations and the exodus to other countries of another million people. Some biotic or abiotic stresses on potato cannot be managed by chemical sprays. Major abiotic stresses (high temperature, drought, soil salinity, and nutrient) and biotic stresses (late blight, bacterial wilt, root-knot nematode and viruses) adversely affect plant development, tuberization, tuber bulking, thus influencing both tuber yield and quality (Minhas, 2012; Wang-Pruski & Schofield, 2012). These factors, along with rapid global population are a major challenge to attain global food security. Moreover, food demand is expected to increase by 59%–98% between 2005 and 2050 (Valin et al., 2014). The growing concerns regarding food security demand crop improvement using diverse approaches for sustainable agriculture.

In recent decades, significant advances in nucleic acid sequencing and information technology have created several omics branches dealing specifically with the molecular components of cellular biology. Genome sequencing provides an unprecedented molecular blueprint for an organism, yielding evolutionarily linked information about their potential metabolic and physiological behavior. Parallel descriptive methods that expand along the canonical flow of biological information (DNA/RNA/proteins/metabolites) by extension from genomics are desired to extend this research beyond the “blueprint” to high-throughput analysis of the molecular changes underlying macroscopic behavior. In the order of this information flow, high-throughput analytical disciplines have emerged, including genomics, transcriptomics, proteomics, metabolomics, ionomics, and phenomics, which together constitute what is known as the “omics cascade” (Fig. 21.1) (Dettmer, Aronov, & Hammock, 2007) which provides a comprehensive view of molecules at the cellular, tissues, or organism level.

A combination of one or more of the “omics platforms” is required to deliver reliable information. Such multiple “omics data sets tend to be very large, as they represent a series of time point and/or treatment samplings and their analysis can only be addressed computationally” (Bieda, 2012; Kohl et al., 2014; Schumacher, Rujan, & Hoefkens, 2014). The prior acquisition of a full genome sequence aids to interpret such data substantially. Combined with phenotyping techniques, the genome sequence of potatoes offers a powerful and rapid tool for determining the genetic basis of agriculturally important traits. The advantage of omics-assisted technology is that genotypic data obtained from a seed or seedling can be used to predict the phenotypic performance of mature individuals without the need of extensive phenotypic evaluation over years and environments. Integrations of various omics methods, techniques, and approaches to advance potato research are discussed in this chapter.

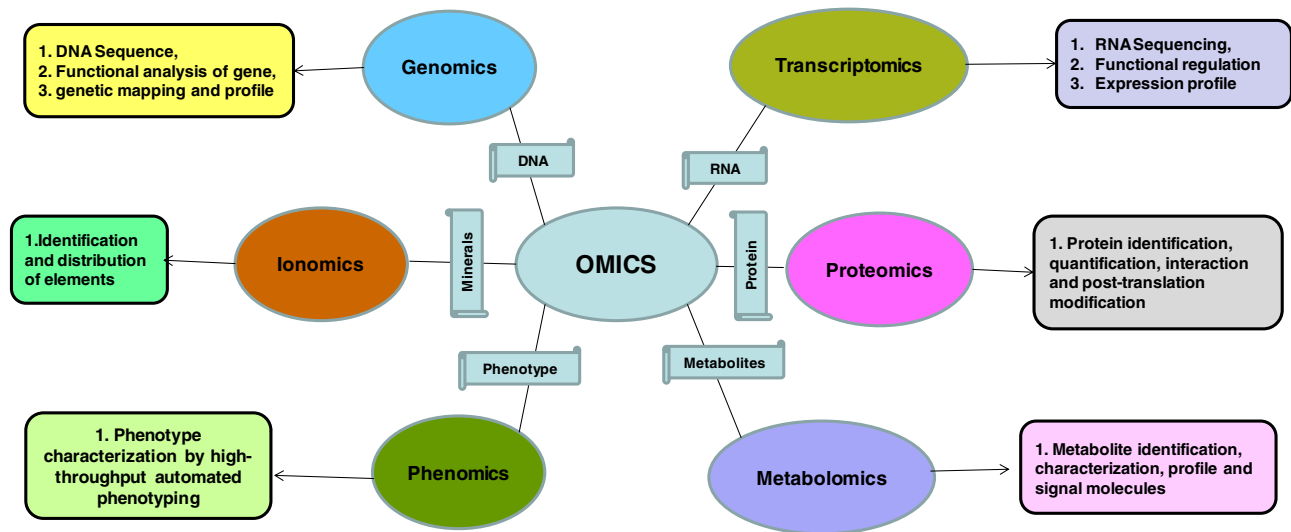


FIGURE 21.1 Different omics-based approaches being used individually or in an integrated manner in plant science.

21.2 Potato genomics

21.2.1 Whole-genome sequencing and resequencing

The genomic understanding of potato has been impeded by a strongly heterozygous and complex genome. The cultivated potatoes are autotetraploid with a basic chromosome number of 12 and an estimated genome size of 840 million base pairs. The availability of the first potato genome sequence and several transcriptomes from a diverse set of potato genotypes, organs, and developmental conditions, have produced genomic tools useful for studying genetic diversity and gene networks that underlie significant characteristics such as disease resistance, tolerance to stress and quality (Kikuchi, Huynh, Endo, & Kazuo, 2015; Ramakrishnan, Ritland, & Blas Sevillano, 2015; Spooner, Ghislain, Reinhard, Jansky, & Gavrilenko, 2014). The publicly accessible potato reference genomes are currently from the doubled monoploid *S. tuberosum* Group Phureja DM1-3 (Potato Genome Sequencing Consortium, 2011), the wild diploid *Solanum commersonii* (Aversano et al., 2015), and the diploid, inbred clone of *Solanum chacoense* M6 (Leisner, Hamilton, Crisovan, Manrique-Carpintero, & Marand, 2018).

Whole-genome resequencing can reveal important differences between cultivated potato varieties and related wild species, especially at a large scale. Traditionally, the cost of resequencing entire populations of samples has been prohibitive, and thus there has been a need for novel solutions to genotype large collections of potato germplasm. Twelve (Kyriakidou, Achakkagari et al., 2020) and six (Kyriakidou, Anglin, Ellis, Tai, & StrÅmvik, 2020) potatoes with different levels of ploidy genomes have recently been sequenced. The two publicly available reference genomes *S. tuberosum* Group Phureja (DM1-3) and *S. chacoense* M6 clone were compared to these genomes for copy number variation (CNV) and single nucleotide polymorphism (SNP) analyses. The study showed the great diversity of potato genomes across this panel and identified a number of CNVs in genes implicated in disease resistance and stress, among other processes.

21.2.2 Molecular markers

Genomic research has a great capability in speeding up breeding processes thus helping in crop improvement like marker-assisted selection and gene pyramiding. In case of potato, several molecular markers have been developed using simple sequence repeat, amplified fragment length polymorphism, nucleotide binding site, or expressed sequence tag (EST) with the aim to genetically localize favorable traits. For example, several studies have identified loci associated with resistance to late blight (Tiwari, Siddappa, & Singh, 2013), potato virus Y (Fulladolsa et al., 2015; Gebhardt, Bellin, & Henselewski, 2006; Nie, Sutherland, & Dickison, 2016; Song, Hepting, & Schweizer, 2005), potato virus X (Gebhardt et al., 2006; Ritter, Debener, & Barone, 1991), and *Verticillium* wilt (Simko, Haynes, & Ewing, 2004; Uribe, Jansky, & Halterman, 2014). There are considerably fewer studies, on the other hand, focusing on polygenic (i.e., quantitative) characteristics such as tuber quality (Li, Tacke, & Hofferbert, 2013) and tuber starch and yield (Schönhals,

Ortega, & Barandalla, 2016). Regardless of trait, many of these low-throughput, gel-based, markers in their current form are not suitable for large-scale screening of progenies, which would be required for application in a breeding program. Recent advancements and tremendous cost reductions associated with high-throughput sequencing have made it possible for different species, including potato the development of genetic markers with a single-nucleotide resolution that can be rapidly assayed on hundreds to thousands of individuals. These molecular markers can be used in applications such as marker-assisted breeding, quantitative trait loci (QTL) determination, genome-wide association studies (GWAS), evolutionary and diversity studies (Uitdewilligen, Wolters, & D'hoop, 2013).

The most popular method in the last decade for high-throughput SNP genotyping has been genotyping arrays. Several SNP arrays for potato have been developed. The Infinium 8303 Potato Array (Felcher, Coombs, & Massa, 2012), which was developed using SNPs identified in two previous studies, is currently one of the most popular: one that mined markers from potato EST databases (Anithakumari, Tang, & van Eck, 2010) and a second that analyzed cDNA sequences from six accessions of elite potato germplasm (Hamilton, Hansey, & Whitty, 2011). The SolSTW array is another recently developed SNP platform. It includes a total of 14,530 SNP markers, most of which were selected from a previous sequence-based genotyping experiment (Vos, Uitdewilligen, & Voorrips, 2015). The design of this array was focused on expanding the genetic sources of the markers, reducing biases and making it more useful for applications such as marker-assisted breeding. Both techniques have become popular in the agricultural genomics and ecological genetics communities, respectively. Among the several SNP-based genotyping methods, the genotyping by sequencing (GBS) approach is a highly multiplexed system for constructing RRL (reduced representation libraries), molecular marker discovery, and genotyping for crop improvement (Elbasyoni et al., 2018; Eltahir et al., 2018). GBS has been applied to many crop species due to low cost and advanced technologies (Kim et al., 2016; Poland & Rife, 2012). In potato, there has been limited application of GBS for molecular marker development perhaps due to the highly heterozygous, tetraploid genome. In one instance, however, a modified GBS approach has been successfully used in marker discovery as part of a study that genotyped a panel of 83 tetraploid potato varieties chosen to represent the most important commercial cultivars and landraces worldwide (Ritter et al., 1991).

21.2.3 Quantitative trait loci mapping, bulked segregant analysis, and GWAS

Marker-based approaches that include genetic fingerprinting, linkage maps, and QTL mapping require extensive genotype data. Linkage mapping and association mapping have led to the detection of QTL by identifying marker–trait associations (Cockram & Mackay, 2018). Much attention has been given to mapping QTLs for many abiotic stresses such as salinity, drought, and low temperatures in potato; however, it is still important to explore other stresses such as elevated temperatures, minimal nutritional regimes, and environmental pollutants (heavy metals, ozone). QTL mapping was performed using a linkage map for drought tolerance in potato, a total of 23 QTLs were identified from control, polyethylene glycol (PEG) stress, and recovery treatments under in vitro conditions. Among these, 10 QTLs were located on chromosome 2, and on linkage groups 2, 3, and 8, 3 QTLs involved in root-to-shoot ratio characteristics were found. In another study by the same group, a total of 47 QTLs were identified in a diploid potato mapping population under well-watered, drought, and recovery conditions (Anithakumari et al., 2010). Among them, 28 QTLs were drought-specific, 17 were specific to the recovery treatment, and 2 were unique to the well-watered condition. A total of 31 significant QTLs were located on chromosomes 5 and 4 for different traits in drought, recovery, and well-watered conditions. Four QTLs for $\delta^{13}C$, three for chlorophyll content, and one for chlorophyll fluorescence (Fv/Fm) were detected to colocalize with yield and other growth trait QTLs identified on other chromosomes. In addition, many QTLs governing stress tolerance, and quality traits have been identified in potato (Table 21.1).

In contrast, a GWAS approach has an advantage over linkage mapping as it explores the genetic diversity and recombination events present in germplasm collections and provides higher mapping resolution (Fukushima, Kusano, Redestig, Arita, & Saito, 2009). Recently, GWAS has been used in potato germplasm collection to detect SNPs for protein content (Klaassen et al., 2019). For instance, GWAS was performed using the SolSTW 20K Infinium SNP marker array where four QTLs have been identified enabling breeding for protein content in potato.

Bulked segregant analysis (BSA) is emerging as a method for genetic mapping that has a particularly good compatibility with genome resequencing. BSA is an approach for gene mapping where pooled DNA from individuals is genotyped as a single bulked sample. The method was originally applied in lettuce using individuals from a single biparental cross that segregated for a downy mildew resistance (Michelmore, Reyes Chin-Wo, & Kozik, 2016), but it can also be used for three-way, four-way, and multiparental crosses, including those developed with special designs such as diallel design, North Carolina design, multiparent advanced generation intercross, and nested association mapping (Zou, Wang, & Xu, 2016). BSA has been used successfully in potato to map steroidal glycoalkaloid content in tetraploids

TABLE 21.1 Significant quantitative trait loci (QTL) mapping studies performed to identify loci governing biotic, abiotic stress tolerance and quality in potato.

Trait	QTL	Chromosome	Position (cM)	LOD score	R%	References
Late blight	qrAUDPC-3	3	68	2.9	3.9	Santa et al. (2018)
	qrAUDPC-8	8	58	3.5	5.2	
	qrAUDPC-5	5	7	2.9	4.2	
	qrAUDPC-4	4	67	2.5	5.4	
	qrAUDPC-3.1	3.1	40	3.1	5.6	
	qrAUDPC-1	1	98	3.4	7.2	
Common Scab	qCS-11	11	41.2	4	18.2	Braun, Endelman, Haynes, and Jansky (2017)
Bacterial wilt	qBWR-1	1	79.1	4.09	11	Habe, Miyatake, Nunome, Yamasaki, and Hayashi (2019)
	qBWR-2	3	15.0	5.56	15.6	
	qBWR-3	7	25.3	5.33	18.4	
	qBWR-4	10	8.8	5.54	15.5	
	qBWR-53	11	35.0	3.26	9.3	
Nematode	qGpaM1	5	4.3	20.9	56	Caromel et al. (2013)
	qGpaM2	6	31	3.7	19	
	qGpaM3	12	48	3.5	15	
Osmotic tolerant	qOS-12	12	-	2.4	26.8	Gorji et al. (2012)
Salt tolerant	qNA-7 (Leaf)	7		4.5	20.6	Losifidis (2011)
	qMg-3 (Stem)	3		7.1	29.4	
	qK/Na-3—(Leaf)	3		4.8	22.4	
	qCl-7(leaf)	7		4.3	19.6	
	qK/Na-1(root)	1		4.2	19.7	
Starch content	qSC-5	5	62.5	-	26.6	Li et al. (2019)
Tuber shape	qTS-1	1	51.5	4.16	12.1	Hara-Skrzypiec, Śliwka, Jakuczun, and Zimnoch-Guzowska (2018)
Eye depth	qED-1	1	69.0	3.74	10.9	
Tuber weight	qTW-1	1	68.9	6.97	19.4	
Tuber flesh color	qTC-2	2	76.8	3.64	10.6	
Cold-induced sweetening	qCIS-4-6	4 and 6	22 and 18	5.4 and 6.0	17.1 and 19.4	Braun et al. (2017)
Starch-corrected chip color	qSCAH-3	3	111.6	3.23	10	Soltys-Kalina et al. (2020)
amylose	qAmyl-2	2	—	5.4	25.1	Acharjee et al. (2018)
Starch gravity	qSG-1	1		4.4	19.3	

(Kaminski, Kørup, & Andersen, 2016). The use of whole-genome sequencing for genotyping will become more common as sequencing costs decrease.

21.3 Potato transcriptomic

Transcriptomics is the field of molecular biology that studies the transcriptome: the complete set of transcripts in a cell, tissue, or organism, which includes the messenger RNA (mRNA) and noncoding RNA (ncRNA) molecules (Morozova, Hirst, & Marra, 2009). To date, several transcriptomic studies have been documented in potato governing response to

abiotic/biotic stress and quality traits (Table 21.2) (Bachem et al., 2000; Barry et al., 2005; Evers et al., 2010; Kloosterman et al., 2008; Massa et al., 2011; Tai et al., 2020). In one of the earliest transcriptomic studies of potato in response to abiotic stresses, 20,756 ESTs from a complementary DNA (cDNA) library were constructed by pooling messenger ribonucleic acid (mRNA) from heat-, cold-, salt-, and drought-stressed potato leaves and roots (Rensink et al., 2005a). Later, to intensify potato transcriptomic analysis, the Potato Oligo Consortium (POCI) array was formed with 44,000 probes representing 42,034 potato unigenes (Kloosterman et al., 2008). The array was integrated into the functional genomics program of a Canadian consortium to improve disease resistance and tuber quality traits of potato (Regan et al., 2006). Recent developments in high-throughput sequencing technologies of the whole transcriptome, known as RNA-Seq, permit the analysis of all transcripts in a sample for mRNA (Fig. 21.2; Table 21.2) and miRNA abundance (Table 21.3) in potato.

TABLE 21.2 Major transcriptomic analysis for biotic, abiotic stress tolerance and quality in potato.

Stress/condition	Platform	DEG	Outcome	References
Late blight	Illumina HiSeqX10	3354 genes	Identified late blight resistance genes	Yang et al. (2018)
Potato virus Y	Illumina HiSeq2000 lanes	407 genes	Identified different Potato virus Y resistance genes	Goyer et al. (2015)
Bacterial wilt	Illumina-Solexa Genome Analyzer II	2978 genes	Identified different bacterial wilt responsible genes in <i>Solanum Commersonii</i>	Zuluaga et al. (2015)
Late blight, bacterial wilt, and necrotic ringspot	Illumina HiSeqTM2500	6945 genes	Insights into the relationship between transcriptome changes infected with the three pathogens	Cao et al. (2020)
Salt stress	Illumina NextSeq	5508	Establish a basis for breeding salt-tolerant cultivars.	Li, Qin, & Hu (2020)
Drought stress	Illumina NextSeq	5118	Transcriptome profiling and characterization of drought-tolerant potato plant	Moon et al. (2018)
	RNA-Seq	1849	Identified different drought responsible genes	Sprenger et al. (2016)
Heat stress	Illumina HiSeq2000	1420	Identified heat stress-tolerant proteins StHsp26-CP and StHsp70	Tang et al. (2020)
	Illumina HiSeq2000	2190 (leaves), 2886 (tubers)	Identified genes associated with ABA, ethylene, auxin, and brassinosteroid, heat-shock proteins, and transcription factors	Hancock et al. (2014)
	Microarray	2500	Identified genes associated with photosynthesis, hormonal activity, sugar transporters, and transcription factors	Hancock et al. (2014)
Low temperature	Microarray	53 groups of putative orthologous genes	Identified key genes responsible for cold tolerant	Carvalho et al. (2012)
Cold, heat, and salt stress	Illumina NextSeq	2584 (cold) 1149 (salt) 998 (heat)	Identified several transcription factors, DNA-binding proteins, transporter proteins, phosphatases, and HSPs response to abiotic stresses	Rensink et al. (2005b)
Nitrogen stress	Illumina NextSeq	Shoots (1041) Stolons (918) Roots (864)	Identified different nitrogen-use efficiency genes	Tiwari et al. (2020a)

DEG, Differentially expressed genes; ABA, abscisic acid; HSP, heat shock protein.

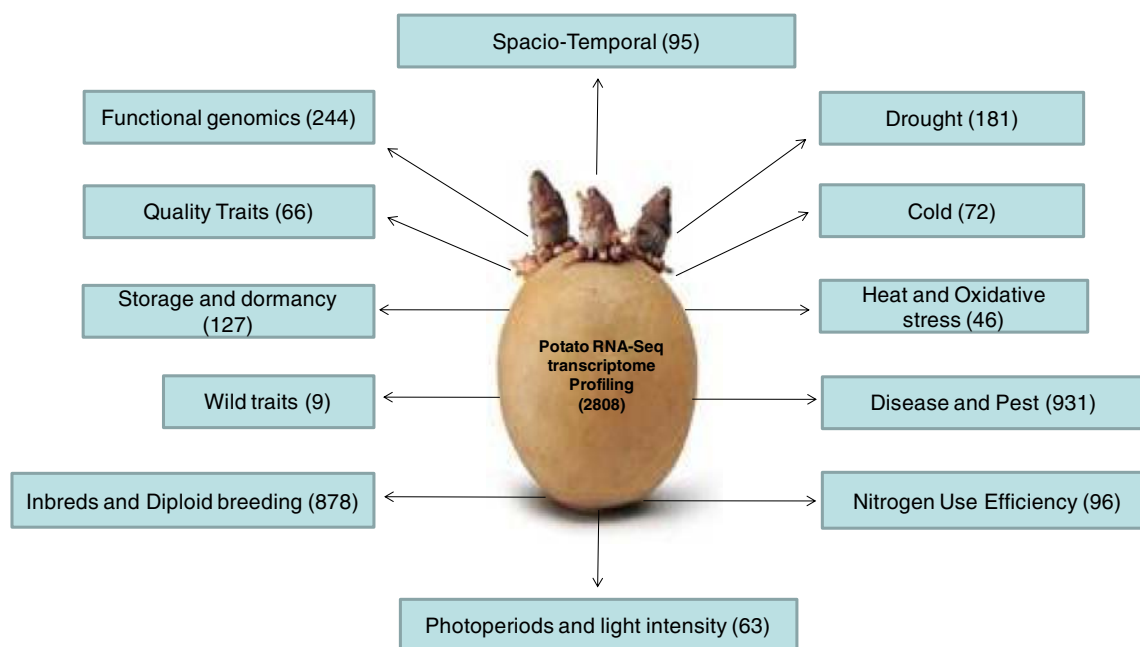


FIGURE 21.2 Transcriptomic resources generated through RNA-Seq approaches in potato being used for different studies. The values provided in the parenthesis indicate approximate number of RNA-Seq libraries publicly available at SRA database (<http://www.ncbi.nlm.nih.gov/sra>).

TABLE 21.3 Potato miRNAs and their biological functions.

miRNAs	Target genes	Biological functions	References
miR8788	<i>StLL1</i>	Late blight resistance	Hu et al. (2020)
miR48, miR397	PR genes	Potato virus A	Li et al. (2017)
miR100	Cytochrome P450	Colorado potato beetles	Mathieu, Morin, Lyons, Sébastien, and Pier (2017)
miR165/166 and 159	Transcription factor gene	Potato virus X	Zhao et al. (2016)
miR162, 168A, and 482	<i>DCL1</i> , <i>AGO1-2</i> and <i>Cc-nbs-lrr</i>	Potato virus Y	Szajko, Yin, and Marczewski (2019)
miR811, miR814, miR835, miR4398	MYB transcription factors	Drought stress	Zhang et al. (2014)
miR166, miR159	Transcription factors	Salinity stress	Kitazumi, Kawahara, Onda, De Koeeyer, and de los Reyes (2015)
miR156, miR169	Transcription factors	Low temperatures stress	Esposito et al. (2020)
miR828	Transcription factors	Purple tuber skin and flesh color	Bonar et al. (2018)
miR397, miR398	—	Nitrogen-use efficiency	Tiwari et al. (2020a)

21.3.1 Biotic stress

Unlike microarray, the RNA-Seq approach is not only confined to compare the transcripts levels, but it is also useful in novel gene discovery and spliced forms, especially in nonmodel plants. During the early stages of potato virus Y infection, RNA-Seq analysis revealed 407 differential upregulated genes in Premier Russet potato variety which is resistant to the PVY strain O (PVY^O) but susceptible to the potato virus Y strain PVY^{NTN}. The genes identified were predicted

to encode for a putative the ATP-binding cassette (ABC) transporters, an MYC2 transcription factor, a VQ motif-containing protein, a nonspecific lipid-transfer protein, and a xyloglucan endotransglucosylase–hydroxylase (Goyer, Hamlin, Crosslin, Buchanan, & Chang, 2015). In another literature, 265 differential expressed genes that are virus-specific and typical potato responses were revealed for potato viruses such as potato virus A (PVA), potato virus Y (PVY), potato leaf roll virus (PLRV) (Osmani et al., 2019).

In a recent study, in potato genotype SD20, 3354 differentially expressed genes (DEGs) were identified, which mainly encoded transcription factors and protein kinases, and also included four NBS-LRR genes. A total of 43 DEGs have been involved in immune response, 19 of which have been enriched by hypersensitive response reactions, which may play an important role in *P. infestans* infection broad-spectrum resistance (Yang et al., 2018). Bali et al. (2019) reported transcriptional changes among *Meloidogyne chitwoodi* resistant and susceptible potato genotypes after nematode inoculation. Differential gene expression analysis reveals that at 48 h, 7 days, 14 days, and 21 days after inoculation, 1268, 1261, 1102, and 2753 genes were upregulated in PA99N82-4, respectively, of which 61 genes were common across all the time points.

21.3.2 Abiotic stress

The impact of drought stress on gene expression has been analyzed with high-throughput transcriptomics in various plants such as rice, maize, and poplar. In a study of potato under drought stress, the transcripts that are differentially expressed under water with holding and rewatering were identified to deepen the understanding of the molecular mechanism of potato stolon responding to water stimulus (Barra et al., 2019; Gong et al., 2015). By analyzing the RNA-Seq data generated from stolon tips of potato plants in each of the three groups, the researchers identified 3189 and 1797 differentially expressed transcripts under only drought treatment and treatment of drought followed by rewatering, respectively. Several of these genes are homologs of known drought-responsive genes in Arabidopsis, including a dehydrin, protein phosphatase, auxin-responsive protein, gibberellic acid–stimulated gene, calmodulin-like protein, abscisic acid 8'-hydroxylases, and calcium-transporting ATPase.

In one of the studies on the heat response of potato at the molecular level, 2190 genes were found to be differentially expressed in potato leaves when the plants were exposed to moderately elevated temperatures (30°C/20°C, day/night) for up to 5 weeks (Hancock et al., 2014). Heat-responsive genes involved in photosynthesis, lipid metabolism, and amino acid biosynthesis were highly overrepresented at all-time points of stress treatment. In tubers a total of 2886 genes exhibited major changes in their transcript levels associated with the different temperature conditions in the course of stress treatment. DEGs in potato tubers were underrepresented in functional categories related to cell wall processes, lipid metabolism, aspects of secondary metabolism, hormone metabolism, biotic stress, DNA metabolism, and development, whereas genes involved in RNA metabolism were overrepresented following moderately high-temperature treatment. In k-means clustering of heat-responsive transcripts of potato, genes associated with ABA, ethylene, auxin, and brassinosteroid responses; heat-shock proteins and transcription factors; and genes previously associated with abiotic stress responses were identified. These data indicate that the potato plants respond differently than other crops to moderately elevated temperatures in such a way that they show a combination of different biochemical and molecular pathways during tuber growth rather than known symptoms of abiotic stress.

To classify abiotic stress-responsive genes in potatoes, Rensink et al. (2005b) subjected potato seedlings to cold (4°C), heat (35°C), or salt (100 mM NaCl) stress for up to 27 h. In at least one stress condition, a total of 3314 cDNA clones exhibited a significant differential expression. A number of 1149 and 998 clones were up- or downregulated under salt and heat stresses, respectively, while 2584 cDNA clones were differentially expressed under cold stress. The functional annotation analysis of differentially expressed clones showed several transcription factors, DNA-binding proteins, transporter proteins, phosphatases, and HSPs (heat shock protein) in response to abiotic stresses.

21.3.3 Quality traits

Cold-responded transcriptome of genotypes between cold-induced sweetening resistant (CIS-R) and cold-induced sweetening sensitive (CIS-S) in tubers was reported by Liu et al. (2020). Comparative transcriptome revealed that in both genotypes, activating the pathways of starch degradation, sucrose synthesis, and hydrolysis could be common strategies in response to cold. Moreover, the variation in sugar accumulation between genotypes may be due to genetic differences in cold response, which could be mainly explained: CIS-R genotype was active in starch synthesis and attenuated in sucrose hydrolysis by promoting the coordinate expression of a series of genes involved in the regulation of the CIS resistance. Liu et al. (2015) explained the molecular mechanism of white and purple potato development, by

identifying differential responses of biosynthetic gene family members together with the variation in structural genes [anthocyanidin 3-*O*-glucosyltransferase (UFGT)] and transcription factors (MYB AN1 and bHLH1) in this highly heterozygous crop.

21.3.4 miRNAs in potato

MicroRNAs (miRNAs) are small (21–24 nt), endogenous, nonprotein-coding RNAs that play important gene regulatory roles in animals and plants by pairing to the mRNAs of protein-coding genes to mediate posttranscriptional repression (Challam, Nandhakumar, & Kardile, 2018). Several microRNAs have been known to regulate abiotic and biotic stresses, yield, and nutritional components in crops. As of March 2018, there were 48,885 plant mature miRNAs from 271 plant species registered in the miRBase database (Challam et al., 2018). In silico analyses have predicted the targets of several miRNAs, and some of them were validated in the laboratory experimentally. The first attempt to identify the miRNAs that function in development and their targets in potato was done in 2009 (Zhang, Luo, Gong, Zeng, & Li, 2009). In this study, 48 potential miRNAs were identified in *S. tuberosum* by in silico comparisons of known miRNAs from other plants against potato EST, and nucleotide databases. Several other researches (Lakhotia, Joshi, & Bhardwaj, 2014; Zhang et al., 2013, 2014) followed this first analysis to identify potato miRNAs, in which the number of predicted miRNAs and their possible potato targets increased significantly, indicating that the prediction algorithms used to identify new miRNAs and their targets have improved in recent years.

High-throughput DNA sequencing allowed researchers to identify various miRNA families affected under biotic, abiotic stresses and quality. A total of 458 known and 674 new miRNAs in control samples were identified in a recent comprehensive deep-sequencing miRNA analysis, while 471 known and 566 novel miRNAs were predicted in drought samples (Zhang et al., 2014). The researchers proved that 100 of the known miRNAs were repressed while 99 of them were induced under 20 days of drought stress in entire potato leaves. Moreover, 151 of the novel miRNAs were repressed while 119 of them were induced in drought-treated potato leaves compared to the controls. In addition, based on target prediction, a total of 246 target genes of known miRNAs and 214 target genes of novel miRNAs were identified. Following the transcript abundance, analyses of selected differentially expressed miRNAs and their target mRNAs, miR811, miR814, miR835, and miR4398 were found to play roles in posttranscriptional regulation of drought-related genes in potato. These miRNAs target an MYB transcription factor, a hydroxyproline-rich glycoprotein, an aquaporin, and a WRKY transcription factor, respectively.

21.4 Potato proteomics

Proteomic approach is used to investigate the responses of plants to stresses as well as complexity of biochemical processes (Aghaei & Komatsu, 2013; Ghosh & Xu, 2014; Gong, Hu, & Wang, 2015). Proteomics has the ability of identifying possible candidate genes that can be used for the genetic enhancement of plants against stresses and quality improvement (Barkla, Vera-Estrella, & Raymond, 2016; Cushman & Bohnert, 2000; Rodziewicz, Swarczewicz, Chmielewska, Wojakowska, & Stobiecki, 2014). Different signaling pathways are reported to be activated in response to stresses resulting in a complex regulatory network involving transcription factors, ion homeostasis, antioxidants, hormones, kinase cascades, reactive oxygen species (ROS), and osmolyte synthesis (Suzuki, Rivero, Shulaev, Blumwald, & Mittler, 2014; Yin et al., 2015). Advances in proteomic technologies have widened our genetic and molecular understanding of plant responses under different stresses and quality traits. Several proteomic studies in potato have been described in Table 21.4.

21.4.1 Biotic stress

Infection of plants by viruses interferes with expression and subcellular localization of plant proteins. Potyviruses comprise the largest and most economically damaging group of plant-infecting RNA viruses. In virus-infected cells, at least two potyviral proteins localize to nucleus but reasons remain partly unknown. The nuclear proteome of leaf cells from the diploid potato line after infection with potato virus A (PVA; genus *Potyvirus*; Potyviridae) was analyzed by Rajamaki, Sikorskaite-Gudziuniene, Sarmah, Varjosalo, and Valkonen (2020) and compared the data with that obtained for healthy leaves. Gel-free liquid chromatography coupled to tandem mass spectrometry was used to identify 807 nuclear proteins in the potato line v2-108; of these proteins, 370 were detected in at least two samples of healthy leaves. Sixteen proteins were predominantly found in samples of leaves contaminated with PVA, while 16 other proteins were specific to healthy leaves. The protein Dnajc14 was only detected in healthy leaves, while the nuclei of PVA-infected

TABLE 21.4 Major proteomic analysis for quality, biotic and abiotic stress tolerance in potato.

Particular	Technique	Number of proteins identified and quantified	Potential for translational research	References
Potato virus A	LC-MS/MS	807 nuclear proteins	Identified PVA infection alters ribosomes and splicing-related proteins in the nucleus of potato leaves	Rajamaki et al. (2020)
Late blight	Quantitative proteomics	8 proteins obtained from leaves	Indicated that CurdOs exhibit activation effect on the early- and late-defense responses in potato leaves	Li et al. (2014)
	Quantitative proteomics	4000 unique proteins	Detect and quantify between 3248 and 3529 unique proteins from each cultivar, and up to 758 <i>Phytophthora infestans</i> –derived proteins	Larsen, Guldstrand, Malene, Bennike, and Stensballe (2016)
Bacterial wilt	2-DE MS	8 proteins obtained from roots	Identified key proteins for bacterial wilt	Ghosh et al. (2016)
Drought stress Drought stress	2-DE-IEF/SDS-PAGE-MALDI-TOF-MS/MS	100 proteins from shoot tip	Identified drought stress tolerant proteins	Bündig, Jozefowicz, Mock, and Winkelmann (2016)
	2-DE/MALDI-TOF-TOF/MS	12 proteins from leaves (Ninglang 182) of potato		Zhang et al. (2013)
Cold stress	2-DE/MALDI-TOF-TOF-MS	94 proteins from shoots of <i>Solanum commersonii</i>	Identified cold stress-tolerant proteins	Folgado et al. (2013)
	2-DE/MALDI-TOF-MS	199 proteins		Folgado et al. (2014)
Cold-induced sweetening tuber	High pH reverse-phase LC (off-gel electrophoresis) + nanoLC–MS/MS (Quantitation: iTRAQ)	46 proteins from tubers	Identification cold-induced sweetening proteins	Yang et al. (2011)
Tuberization process	Shotgun proteomic approach	251 proteins	Whole tuberization proteome profiling	Lehesranta et al. (2006), Yu et al. (2012)

leaves were overrepresented by different ribosomal proteins, ribosome biogenesis proteins, and RNA splicing-related proteins. Two virus-encoded proteins were identified in the samples of PVA-infected leaves. The data indicate that potyvirus infection particularly affects ribosomes and splicing-related proteins in the nucleus.

The proteome dynamics of potato cv. Sarpo Mira was studied by Xiao et al. (2019) after foliar application of zoospore suspension from *P. infestans* isolate, at three key time points: zero hours post inoculation (hpi) (control), 48 hpi (EI), and 120 hpi (LI); divided into early and late disease stages by the tandem mass tagging method. A total of 1229 differentially expressed proteins (DEPs) were identified in cv. Sarpo Mira in a pairwise comparison of the two disease stages, including commonly shared DEPs, specific DEPs in early and late disease stages, respectively. In the early stages of infection, over 80% of the protein abundance changes were upregulated, while more DEPs (61%) were down-regulated in the later stage of the disease. Expression patterns, functional category, and enrichment tests highlighted significant coordination and enrichment of cell wall-associated defense response proteins during the early stage of

infection. The late stage was characterized by a cellular protein modification process, membrane protein complex formation, and cell death induction.

21.4.2 Abiotic stress

Two-dimensional (2D) gel electrophoresis was used to analyze potato treated with salt stress by applying 90-mM NaCl. A total of 322 and 305 differentially expressed proteins were detected in shoots of Kennebec and Concord, respectively. These proteins differentially expressed under NaCl treatment were involved in protein synthesis, metabolism/energy, and photosynthesis. Markedly upregulated were osmotin-like proteins, TSI-1 protein, heat-shock proteins, protein inhibitors, calreticulin proteins (Aghaei, Ehsanpour, & Komatsu, 2008). The osmotin-like proteins, TSI-1 protein, heat-shock proteins, protein inhibitors, calreticulin proteins were markedly upregulated (Aghaei et al., 2008). In another study, potato responses to salt (150-mM NaCl) stress were investigated using proteomic approach by Evers et al. (2012). A substantial amount of protein abundances was found significantly changed under salt treatment. A strong decrease in the abundance of photosynthesis-related proteins was caused by salt exposure. The proteins involved in primary metabolism such as glyceraldehyde-3-phosphate dehydrogenase, triosephosphate isomerase were strongly repressed. Nitrogen and amino acid metabolisms related proteins were also decreased after salt treatment, especially polyamine synthesis-related proteins, such as arginine decarboxylase, *S*-adenosylmethionine decarboxylase, agmatine deiminase (Evers et al., 2012). Protein analysis detected by proteomics in salt-stressed vegetable crops may help further elucidate salt stress resistance and protection mechanisms in higher plants.

A comparative study with differences in the protein group analysis of the potato drought resistance variety in Ninglang 182 leaves was investigated by Zhang et al. (2013), using 2D gel electrophoresis during drought and normal processing conditions. There were 12 differentially expressed protein spots identified by electrophoresis and MALDI-TOF-TOF/Ms analysis were drought resistance proteins of potato variety Ninglang 182.

21.4.3 Quality traits

Blue-native PAGE and 2-DE were used to identify and characterize mitochondrial protein complexes from various plants, including potato tubers (Eubel, Heinemeyer, & Braun, 2004). A total of 18 proteins were identified from several supercomplexes and respiratory complexes using Arabidopsis and bean as model systems (Eubel et al., 2004). Based on their findings, authors concluded that supercomplex formation between complexes I and III reduces the access of alternative oxidase to its substrate and probably regulates alternative respiration (Eubel, Heinemeyer, & Braun, 2003). In another study, using shotgun proteomics approach containing 1060 nonredundant proteins, potato tuber mitochondrial proteome was established. The components of electron transport chain, tricarboxylic acid cycle, and protein import apparatus were the most abundant mitochondrial proteins. Some of the other identified proteins included 71 pentatricopeptide repeat proteins, 29 membrane carriers/transporters, proteins involved in coenzyme biosynthesis and iron metabolism, pyruvate dehydrogenase kinase, and a type 2C protein phosphatase. In addition, the presence of PTM sites was demonstrated by N50% of the identified proteins, suggesting a vast regulation of mitochondrial proteins at posttranslational level (Salvato et al., 2014). By colocalization on the genetic map and a direct correlation study of protein abundances and phenotypic traits, a relationship between proteins and 26 potato tuber quality traits (e.g., flesh color, enzymatic discoloration) was established (Acharjee et al., 2018). Over 1643 unique protein spots were detected in total over the two harvests. For over 300 different protein spots, they were able to map pQTLs, some of which were colocalized with characteristics such as starch content and cold sweetening.

21.5 Potato metabolomics

Metabolomics refers to the quantification and identification of metabolites present in the biological organization such as cells, tissues, organs, biofluids, or whole organisms at a certain period of time (Daviss, 2005). Metabolites are the end products of metabolism; more precisely any molecule, the size of which is less than 1 kDa, comes under this category (Samuelsson & Larsson, 2008). Metabolomics is downstream of transcriptomics and proteomics. Unlike two others, the size of metabolome of a species cannot be hypothesized by tools that use existing genomic information on central dogma principle. Analysis of intricate metabolite interactions for key players of pathways leads to significant understanding of individual genomic information and metabolic outputs (Toubiana, Fernie, Nikoloski, & Fait, 2013).

21.5.1 Biotic traits

Plants defend themselves from pathogens by producing bioactive defense chemicals. Biochemical pathways related to the quantitative resistance of potato to *Spongospora subterranea* f. sp. *subterranea* (Sss) are, however, not understood and are not efficiently utilized in potato breeding programs. Untargeted metabolomics using ultraperformance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry (UPLC-Q-TOF/Ms) was used to elucidate the biochemical mechanisms of susceptibility to Sss root infection. To identify tolerance-related metabolites, potato roots and root exudate metabolic profiles of five tolerant cultivars were compared with those of five susceptible cultivars following Sss inoculation. Contrasting responses to Sss infection were exposed when comparing the relative metabolite abundance of resistant versus susceptible cultivars. Metabolites belonging to amino acids, organic acids, fatty acids, phenolics, and sugars, as well as well-known cell wall thickening compounds were putatively identified and were especially abundant in the tolerant cultivars relative to the susceptible cultivars. Compared to susceptible cultivars following Sss inoculation, metabolites known to activate plant secondary defense metabolism were significantly increased in the tolerant cultivars. Root-exuded compounds belonging to the chemical class of phenolics were also found in abundance in the tolerant cultivars compared to susceptible cultivars (Lekota, Modisane, Apostolides, & van der Waals, 2020).

21.5.2 Abiotic traits

The most promising method to decipher abiotic stress tolerance in plant species has emerged from metabolomics. Recently, metabolomics has been applied to probe for unique metabolites during the life cycle of plants. Biotic/abiotic stresses have a significant role in the reduction of the crop yield (Hein, Sherrard, Manfredi, & Abebe, 2016). The importance of potato is difficult to overestimate; it is a valuable source of carbohydrates, antioxidants, and vitamins. A large number of investigations are focused on the study of metabolic processes occurring in the potato plant to elucidate the mechanisms responsible for productivity, accumulation of the compounds that determine taste and nutritional quality, maintaining the quality of tubers in storage, plant resistance to pathogens, etc. The sum of the metabolites generated as a consequence of the activity of the metabolic network is known as the metabolome. Complex studies of metabolic diversity with the use of modern state-of-the-art chromatography approaches and the highly precise detection of individual compounds revealed the specificity of metabolic spectra from the subcellular to the organismal levels and its amazing plasticity under the influence of a variety of internal and external stimuli. Metabolomic approaches are already in use for phenotyping the available species, lines, and varieties, as well as for assessing the tolerance of potato plants to environmental challenges and for detecting changes in tubers during storage (Puzanskiy, Yemelyanov, Gavrilenko, & Shishova, 2017). The use of metabolomics to research biotic/abiotic stress will help us to elucidate underlying molecular mechanisms associated with stress and would surely lead to developing tolerant potato plants with enhanced yield.

21.5.3 Quality traits

Potato contains phytochemicals with demonstrated effects on human health was reported by Chaparro, Holm, Broeckling, Prenni, and Heuberger (2018). A comprehensive metabolomics (UPLC- and GC—Ms) and ionomics (ICP—Ms) analysis of raw and cooked potato tuber was performed on 60 unique potato genotypes that span 5 market classes, including russet, red, yellow, chip, and specialty potatoes. A total of 2656 compounds comprising known bioactives (43 compounds), nutrients (42), lipids (76), and 23 metals were identified in the study. Most nutrients and bioactives were partially degraded during cooking (44 out of 85; 52%); however, genotypes with high quantities of bioactives remained highest in the cooked tuber. Chemical variation was influenced by genotype and market class. In particular, ~53% of all detected compounds from cooked potato varied among market class and 40% varied by genotype. The most prominent metabolite profiles were observed in yellow-flesh potato which had higher levels of carotenoids and specialty potatoes which had higher levels of chlorogenic acid as compared to the other market classes. In addition, more metabolite variance was found within the market class (e.g., α -tocopherol, ~onefold variation among market class vs ~threefold variation within market class). Taken together, the study characterized significant metabolite and mineral variation in raw and cooked potato tuber and supports the potential to breed new cultivars for improved health traits.

The assessment of unintended impacts on plant-insect interactions is an important aspect of ecological protection for genetically modified (GM) plants. The chemical composition of plants is determined to a large degree by these interactions. This study uses nuclear magnetic resonance (NMR)-based metabolomics to establish a baseline of chemical variation to which differences between a GM potato line and its parent cultivar are compared. The effects of leafage, virus infection, and aphid herbivory on plant metabolomes were studied. Only in young leaves of noninfected plants did the

metabolome of the GM line differ from its parent. This effect was small when compared to the baseline. Consistently, aphid performance on excised leaves was influenced by leafage, while no difference in performance was found between GM and non-GM plants. The metabolomic baseline approach is concluded to be a useful tool in ecological safety assessment explained by [Plischke, Choi, Brakefield, Klinkhamer, and Bruinsma \(2012\)](#).

21.6 Potato ionomics

With the blending of concepts from both metabolomics and plant mineral nutrition, the inception of ionomics occurred. Lahner and colleagues first described the ionome to include all the metals, metalloids, and nonmetals present in an organism ([Lahner, Gong, Mahmoudian, Smith, & Abid, 2003](#)), expanding the term metallome to include biologically significant nonmetals such as nitrogen, phosphorus, sulfur, selenium, chlorine, and iodine ([Outten & O'Halloran, 2001](#); [Williams, 2001](#)). It is important to note here that there are blurred borders between the ionome, the metabolome, and the proteome. For example, phosphorus, sulfur, or nitrogen compounds containing nonmetals would fall within both the ionome and the metabolome, and metals such as zinc, copper, manganese, and iron would fall within the proteome or metalloproteome as defined ([Szpunar, 2004](#)). Ionomics has the advantage in revealing network among various mineral elements in organism ([Baxter et al., 2008](#)). For example, ionic analyses have been performed to isolate the genes that are responsible for mineral transport and homeostasis in plants ([Chao et al., 2011](#)). To examine phylogenetic and environmental effects on plant mineral accumulation, ionomics may also be used ([Sha et al., 2012](#)). A wide range of studies have been done in the field of ionomics mainly on silicon. Most of the dicots and particularly the Solanaceae family take up small quantities of silicon and accumulate less than 0.5% in their tissue. Silicon has been found to enhance drought tolerance and delay in wilting and benefit certain plants when they are under stress.

21.7 Phenomics

The major aim of genetics is to understand phenotypic characteristics and their variation developed through a complex network of interactions between genetic and environmental factors. As a result of their contact with their environment in the ecosystem, the phenotype is the manifestation of a genotype. The characterization of phenotype in multiple levels considering various environmental and external factors affecting the phenotype collectively results in phenomics ([Araus, Kefauver, Zaman-Allah, Olsen, & Cairns, 2018](#); [Dhondt, Wuyts, & Inze, 2013](#)). Phenomics is the translation of genes or the whole genome into the phenotype of plants through recent advances in genomics, and analysis of large datasets relating to the traits under consideration. To understand a genotype and to plan breeding and genetic studies for crop enhancement, integration of this knowledge is important. However, characterization of phenomes lags much behind the developments in the area to characterize genomes.

With the rapid development of novel sensors, imaging technology, and analysis methods, numerous infrastructure platforms have been developed for phenotyping. Over the last two decades the rapid development of nondestructive sensing and imaging techniques has dramatically advanced the measurement of crop phenotypic traits in controlled environments as well as in the field ([Milella, Marani, Petitti, & Reina, 2019](#)). The imaging techniques include visible, thermal infrared, fluorescence, 3D, and multi- or hyperspectral imaging and tomographic imaging by magnetic resonance imaging or X-ray computed tomography ([Jin et al., 2021](#)). Integration of sensing technologies, automatic control technology, computers, robotics, and aeronautics has led to the development of an increasing number of high-throughput phenotyping platforms for investigating crop phenotypic traits. Scientists have developed multiple phenotyping platforms for crop traits at multiple application scales; however, their application in potato has so far been very few or limited. In potato, computer vision and machine learning techniques have been used for the recognition of different diseases on different scales from the tissue to the canopy level. Atherton and Watson ([Atherton & Watson, 2015](#); [Atherton, Choudhary, & Watson, 2017](#)) used hyperspectral remote sensing for detection of early blight (*Alternaria solani*) in potato plants prior to visual disease symptoms. For late blight (*P. infestans*) detection, [Ray, Jain, Arora, Chavan, and Panigrahy \(2011\)](#) also used a point spectrum approach without using spatial information. [Hu, Ping, Xu, Shan, and He \(2016\)](#) used hyperspectral imaging to detect late blight disease on potato leaves successfully, with a discrimination of 95% between healthy and diseased leaves. [Mohanty, Hughes, and Salathé \(2016\)](#) suggested using deep learning convolutional neural network methods for disease identification in plants. They used a large set of 54,306 images of 38 classes of different plants and diseases, including potatoes, and reported an accuracy between 85% and 99% on the different classes. Similarly, [Oppenheim, Shani, Erlich, and Tsror \(2019\)](#) applied deep convolutional neural network to construct database of images for the detection of four potato diseases, black scurf, silver scurf, common scab, and black dot. In another study, [Griffel, Delparte, and Edwards \(2018\)](#) adapted a fully convolutional neural network using

hyperspectral imaging and deep learning for the detection of PVY-infected potato plants. Although virus diseases have a different mechanism by which they change the plant physiology, virus symptoms can also be measured using optical techniques. Spectral signatures of potato plants infected with PVY acquired with a handheld device were classified with an accuracy of 89.8% between infected and noninfected plants. In this the precision and recovery exceeded 0.78 and 0.88, respectively, compared with conventional disease assessment.

Because the phenotype is the product of the genotype and its interaction with the environment, replicating the environmental conditions for screening and cataloging any phenotypic variation that occurs in the genotype is very important for precise phenotypic variations to be obtained. In potato, aeroponic culture technique is an optional device of soil-less culture in growth-controlled environments such as greenhouses. Recently, aeroponics-based precision phenotyping enables identification of nitrogen-use efficient (Tiwari, Devi, et al., 2020) and iron-deficient tolerant (Clarissa et al., 2021) genotypes based on key traits and genes involved. Multidimensional, high-resolution data on agronomical, physiological, and morphological traits describing the phenotype in optimal, biotic, and abiotic stress conditions would enable mapping of genetic elements to biological function at the desired level of accuracy. There is a great need for a combination of technology and proper study to reliably estimate phenomes. Because of their high throughput, phenomics studies are resource intensive and will support many studies due to the generation of large-scale data, and the quality of the data depends on the existing genetic diversity, growth conditions used, phenotypic assays, and further data collection, storage, and interpretation.

21.8 Potato omics resources and integration of technologies

A great deal of data has been produced in recent progress in omics (genomics, transcriptomes, proteomics, metabolomics, ionomics, and phenomics), which can be used to identify novel genetic and chemical elements that regulate various physiological processes (Cohen, Aharoni, Szymanski, & Dominguez, 2017). But the complexity of a trait in plants needs convergence of different approaches to understand complex stress response (Fig. 21.1). In addition, the analysis of high-throughput data from various omics approaches is one of the biggest challenges to interpreting the response mechanism(s). As pointed out by Scholz, Gatzek, Sterling, Fiehn, and Selbig (2014), the development of software tools to enable in-depth analysis of any list of interrelated biological data (pathway analysis tools) is evolving.

Although there are several potentially useful applications for gene expression arrays, it is not simply mRNA levels that need to be considered, but also the amount and modification of proteins expressed that determine true gene activity.

TABLE 21.5 Online databases/tools for integrated omics in potato.

Databases/tools	Context in which the database is important for potato and integrative biology	Accessed by
Spud DB	Datasets and data mining tools to view and analyze the potato genome, including tools to facilitate breeding improved cultivars	http://solanaceae.plantbiology.msu.edu
PlantGDB	Sequence download, BLAST analysis, and multiple sequence alignment	http://www.plantgdb.org
Sol Genomics Network	Potato maps and markers, BAC/EST sequences	https://solgenomics.net
PoMaMO	Retrieval of sequence, SNP, mapping data	https://www.gabipd.org/projects/Pomamo/
PotatoCyc	Metabolic pathway prediction	https://plantcyc.org/databases/potatocyc/4.0
PATHWAY	Information on metabolites and genes, as well as graphical representations of metabolic pathways and complexes	https://www.genome.jp/kegg/pathway.html
Mapman4	Enrichment analysis and visualization of data expression	https://mapman.gabipd.org
MixOmic	Data integration and similarity relationship	http://www.mixOmics.org
Paintomics3	Pathway analysis and interaction	http://www.paintomics.org
Pathview	Data integration and visualization	https://pathview.uncc.edu/
PathVisio 3	Pathway editor and data visualization	https://pathvisio.github.io/

Therefore an important goal is to couple transcriptomic data with other omics tools, proteomics, and metabolomics to establish an integrated understanding of biological processes that, for example, regulate the crop plant composition. These “omics” tools in turn must be linked to DNA sequences and sequence variation to better understand the processes which contribute to biological variation. The goal of “omic” approaches is, therefore, to acquire a comprehensive, integrated understanding of biology by studying all biological processes to identify the various players (e.g., genes, RNA, proteins, and metabolites) rather than each of those individually.

Several metabolic pathway databases are available to facilitate our understanding of transcriptome and metabolome data. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG; <http://www.genome.ad.jp/kegg/>) has a pathway database (PATHWAY) that contains information on metabolites and genes, as well as graphical representations of metabolic pathways and complexes derived from various biological processes. Successful combination of various methods, techniques, and approaches is a promising strategy to develop new climate-smart crop varieties (Tiwari, Challam, Chakrabarti, & Feingold, 2020) under various stresses like nitrogen deficiency (Tiwari, Buckseth, Singh, Kumar, & Kant, 2020; Tiwari et al., 2020b; Tiwari, Buckseth, Devi et al., 2020; Tiwari, Plett, Garnett, Chakrabarti, & Singh, 2018). Efficient adaptation of computational techniques by the plant breeder largely depends on features such as user-friendly interface, easy access, online tutorials and manuals, and interactive options. In this regard, a few user-friendly databases that could be useful to integrate omics scale data from different approaches in potato are described in Table 21.5.

21.9 Conclusions

Various omics tools and techniques have been developed to understand the molecular mechanisms underlying plant response. Under all kind of stress conditions, plants modulate themselves to adopt the existing stresses by controlling gene regulation, proteins, and metabolites. It is essential to elucidate the functions of newly identified genes/proteins/metabolites to understand the stress responses of plants. To identify changes various tools and techniques like genomics, transcriptomics, metabolomics, ionomics, and phenomics have been devised to allow the understanding of genetic makeup in depth, their signaling cascade, and their adaptability under stress conditions. Genomics, transcriptomics, proteomics, and metabolics have been established in potato, but the other branches are still lingering behind, such as ionomics and phenomics. A diverse analysis of omics tools and data integration is needed to make sense and relate the data back to the objective of the research to effectively handle biotic/abiotic stresses and quality traits.

References

- Acharjee, A., Chibon, P. Y., Kloosterman, B., America, T., Renaut, J., Maliepaard, C., . . . Visser, R. G. F. (2018). Genetical genomics of quality related traits in potato tubers using proteomics. *BMC Plant Biology*, *18*(1), 20. Available from <https://doi.org/10.1186/s12870-018-1229-1>.
- Aghaei, K., Ehsanpour, A. A., & Komatsu, S. (2008). Proteome analysis of potato under salt stress. *Journal of Proteome Research*, *4*, 4858–4868.
- Aghaei, K., & Komatsu, S. (2013). Crop and medicinal plants proteomics in response to salt stress. *Frontiers in Plant Science*, *4*, 8. Available from <https://doi.org/10.3389/fpls.2013.00008>.
- Anithakumari, A. M., Tang, J., & van Eck, H. J. (2010). A pipeline for high throughput detection and mapping of SNPs from EST databases. *Molecular Breeding*, *26*, 65–75.
- Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., & Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends in Plant Science*, *23*, 451–466.
- Atherton, D., Choudhary, R., & Watson, D. (2017). *Hyperspectral remote sensing for advanced detection of early blight (Alternaria solani) disease in potato (Solanum tuberosum) plants prior to visual disease symptoms*. St. Joseph, MI: American Society of Agricultural and Biological Engineers.
- Atherton, D., & Watson, D. G. (2015). Hyperspectral spectroscopy for detection of early blight (*Alternaria solani*) disease in potato (*Solanum tuberosum*) plants at two different growth stages. In *ASABE Annual International Meeting* (pp. 1660–1674). St. Joseph, MI.
- Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., . . . Carputo, D. (2015). The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *The Plant Cell*, *27*(4), 954–968.
- Bachem, C., van der Hoeven, R., Lucker, J., Ronald, O., Emanuela, C., Evert, J., . . . Richard, V. (2000). Functional genomics analysis of potato tuber life-cycle. *Potato Research*, *43*, 297–312.
- Bali, S., Vining, K., Gleason, C., Majtahedi, H., Brown, C. R., & Sathuvalli, V. (2019). Transcriptome profiling of resistance response to *Meloidogyne chitwoodi* introgressed from wild species *Solanum bulbocastanum* into cultivated potato. *BMC Genomics*, *20*(1), 907. Available from <https://doi.org/10.1186/s12864-019-6257-1>.
- Barkla, B. J., Vera-Estrella, R., & Raymond, C. (2016). Single-cell-type quantitative proteomic and ionic analysis of epidermal bladder cells from the halophyte model plant *Mesembryanthemum crystallinum* to identify salt-responsive proteins. *BMC Plant Biology*, *16*, 110. Available from <https://doi.org/10.1186/s12870-016-0797-1>.

- Barra, M., Meneses, C., Riquelme, S., Pinto, M., Lagüe, M., Davidson, C., ... Tai, H. H. (2019). Transcriptome profiles of contrasting potato (*Solanum tuberosum* L.) genotypes under water stress. *Agronomy*, 9(12), 848.
- Barry, F., Charlotte, R., Rebecca, G., Martin, L., David, D., Ravinder, S., ... Sharon, R. (2005). Potato expressed sequence tag generation and analysis using standard and unique cDNA libraries. *Plant Molecular Biology*, 59(3), 407–433.
- Baxter, I. R., Vitek, O., Lahner, B., Muthukumar, B., Borghi, M., Morrissey, J., ... Salt, D. E. (2008). The leaf ionome as a multivariable system to detect a plant's physiological status. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 12081–12086.
- Bieda, M. (2012). Kepler for 'omics bioinformatics. *Procedia Computer Science*, 9, 1635–1638.
- Birch, P. R., Bryan, G., Fenton, B., Gilroy, E. M., Hein, I., & Jones, J. T. (2012). Crops that feed the world: Potato are the trends of increased global production sustainable. *Food Security*, 477–508.
- Bonar, N., Liney, M., Zhang, R., Austin, C., Dessoly, J., Davidson, D., ... Hornyik, C. (2018). Potato miR828 is associated with purple tuber skin and flesh color. *Frontiers in Plant Science*, 9, 1742. Available from <https://doi.org/10.3389/fpls.2018.01742>.
- Braun, S. R., Endelman, J. B., Haynes, K. G., & Jansky, S. H. (2017). Quantitative trait loci for resistance to common scab and cold-induced sweetening in diploid potato. *The Plant Genome*, 10(3). Available from <https://doi.org/10.3835/plantgenome2016.10.0110>.
- Bündig, C., Jozefowicz, A. M., Mock, H. P., & Winkelmann, T. (2016). Proteomic analysis of two divergently responding potato genotypes (*Solanum tuberosum* L.) following osmotic stress treatment in vitro. *Journal of Proteomics*, 143, 227–241.
- Cao, W., Gan, L., Shang, K., Wang, C., Song, Y., Liu, H., ... Zhu, C. (2020). Global transcriptome analyses reveal the molecular signatures in the early response of potato (*Solanum tuberosum* L.) to *Phytophthora infestans*, *Ralstonia solanacearum*, and Potato virus Y infection. *Planta*, 252(4), 57. Available from <https://doi.org/10.1007/s00425-020-03471-6>.
- Caromel, B., Mugniéry, D., Lefebvre, V., Andrzejewski, S., Ellissèche, D., Kerlan, M. C., ... Rousselle-Bourgeois, F. (2013). Mapping QTLs for resistance against *Globodera pallida* (Stone) Pa2/3 in a diploid potato progeny originating from *Solanum spegazzinii*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 106(8), 1517–1523.
- Carvalho, M. A., Pino, M. T., Jeknić, Z., Zou, C., Doherty, C. J., Shiu, S. H., ... Thomashow, M. F. (2012). A comparison of the low temperature transcriptomes and CBF regulons of three plant species that differ in freezing tolerance: *Solanum commersonii*, *Solanum tuberosum*, and *Arabidopsis thaliana*. *Journal of Experimental Botany*, 62(11), 3807–3819.
- Challam, C., Nandhakumar, N., & Kardile, H. B. (2018). Advances in Crop Improvement: Use of miRNA technologies for crop improvement. In R. Banerjee, G. V. Kumar, & S. P. J. Kumar (Eds.), *OMICS-based approaches in plant biotechnol* (pp. 55–74). Wiley-Scrivener.
- Chao, D. Y., Gable, K., Chen, M., Baxter, I., Dietrich, C. R., Cahoon, E. B., ... Salt, D. E. (2011). Sphingolipids in the root play an important role in regulating the leaf ionome in *Arabidopsis thaliana*. *The Plant Cell*, 23(3), 1061–1081.
- Chaparro, J. M., Holm, D. G., Broeckling, C. D., Prenni, J. E., & Heuberger, A. L. (2018). Metabolomics and ionomics of potato tuber reveals an influence of cultivar and market class on human nutrients and bioactive compounds. *Frontiers in Nutrition*, 5, 36. Available from <https://doi.org/10.3389/fnut.2018.00036>.
- Clarissa, C., Dutt, S., Sharma, J., Bag, T. K., Raveendran, M., & Sudhakar, D. (2021). Screening for iron deficient chlorosis (IDC) tolerant genotypes in potato (*Solanum tuberosum*, L.) under aeroponic system. *Asian Journal of Microbiology, Biotechnology & Environmental Sciences*, 23(1), 92–99.
- Cockram, J., & Mackay, I. J. (2018). Genetic mapping populations for conducting high-resolution trait mapping in plants. In R. K. Varshney, M. K. Pandey, & A. Chitkineni (Eds.), *Plant genetics and molecular biology* (pp. 109–138). Cham, Switzerland: Springer.
- Cohen, H., Aharoni, A., Szymanski, J., & Dominguez, E. (2017). Assimilation of 'omics' strategies to study the cuticle layer and suberin lamellae in plants. *Journal of Experimental Botany*, 68, 5389–5400.
- Cushman, J. C., & Bohnert, H. J. (2000). Genomic approaches to plant stress tolerance. *Current Opinion in Plant Biology*, 2000, 117–124.
- Daviss, B. (2005). Growing pains for metabolomics. *Scientist (Philadelphia, PA.)*, 19, 25–28.
- Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26, 51–78.
- Dhondt, S., Wuyts, N., & Inze, D. (2013). Cell to whole-plant phenotyping: The best is yet to come. *Trends in Plant Science*, 18, 428–439.
- Elbasyoni, I. S., Lorenz, A., Guttieri, M., Frels, K., Baenziger, P., Poland, J., ... Akhunov, E. A. (2018). Comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Science (Shannon, Ireland)*, 270, 123–130.
- Eltaher, S., Sallam, A., Belamkar, V., Emara, H. A., Nower, A. A., Salem, K. F. M., ... Baenziger, P. S. (2018). Genetic diversity and population structure of F3:6 Nebraska winter wheat genotypes using genotyping-by-sequencing. *Frontiers in Genetics*, 9, 76. Available from <https://doi.org/10.3389/fgene.2018.00076>.
- Esposito, S., Aversano, R., Bradeen, J. M., Di Matteo, A., Villano, C., & Carputo, D. (2020). Deep-sequencing of *Solanum commersonii* smallRNA libraries reveals ribo regulators involved in cold stress response. *Plant Biology (Stuttgart, Germany)*, 22(Suppl. 1), 133–142.
- Eubel, H., Heinemeyer, J., & Braun, H. P. (2004). Identification and characterization of respirasomes in potato mitochondria. *Plant Physiology*, 134, 1450–1459.
- Eubel, H. L., Heinemeyer, J., & Braun, H. P. (2003). New insights into the respiratory chain of plant mitochondria. Supercomplexes and a unique composition of complex II. *Plant Physiology*, 133, 274–286.
- Evers, D., Lefevre, I., Legay, S., Lamoureux, D., Hausman, J. F., Rosales, R. O. G., ... Schafleitner, R. (2010). Identification of drought-responsive compounds in potato through a combined transcriptomic and targeted metabolite approach. *Journal of Experimental Botany*, 61(9), 2327–2343.
- Evers, D., Legay, S., Lamoureux, D., Hausman, J. F., Hoffmann, L., & Renaut, J. (2012). Towards a synthetic view of potato cold and salt stress response by transcriptomic and proteomic analyses. *Plant Molecular Biology*, 2012, 503–514.
- Felcher, K. J., Coombs, J. J., & Massa, A. N. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One*, 7, e36347.

- Folgado, R., Panis, B., Sergeant, K., Renaut, J., Swennen, R., & Hausman, J. F. (2013). Differential protein expression in response to abiotic stress in two potato species: *Solanum commersonii* Dun and *Solanum tuberosum* L. *International Journal of Molecular Sciences*, 4912–4933.
- Folgado, R., Sergeant, K., Renaut, J., Swennen, R., Hausman, J. F., & Panis, B. (2014). Changes in sugar content and proteome of potato in response to cold and dehydration stress and their implications for cryopreservation. *Journal of Proteomics*, 98, 99–111.
- Fukushima, A., Kusano, M., Redestig, H., Arita, M., & Saito, K. (2009). Integrated omics approaches in plant systems biology. *Current Opinion in Chemical Biology*, 13, 532–538.
- Fulladolsa, A. C., Navarro, F. M., Kota, R., Severson, K., Palta, J. P., & Charkowski, A. O. (2015). Application of marker assisted selection for Potato Virus Y resistance in the University of Wisconsin Potato Breeding Program. *American Journal of Potato Research*, 92, 444–450.
- Gebhardt, C., Bellin, D., & Henselewski, H. (2006). Marker-assisted combination of major genes for pathogen resistance in potato. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 112, 1458–1464.
- Ghosh, D., & Xu, J. (2014). Abiotic stress responses in plant roots: A proteomics perspective. *Frontiers in Plant Science*. Available from <https://doi.org/10.3389/fpls.2014.00006>.
- Ghosh, S., Narula, K., Sinha, A., Ghosh, R., Jawa, P., Chakraborty, N., ... Chakraborty, S. (2016). Proteometabolomic study of compatible interaction in tomato fruit challenged with *Sclerotinia rolfsii* illustrates novel protein network during disease progression. *Frontiers in Plant Science*, 7, 1034. Available from <https://doi.org/10.3389/fpls.2016.01034>.
- Gong, F., Hu, X., & Wang, W. (2015). Proteomic analysis of crop plants under abiotic stress conditions: Where to focus our research? *Frontiers in Plant Science*, 6, 418. Available from <https://doi.org/10.3389/fpls.2015.00418>.
- Gong, L., Zhang, H., Gan, X., Zhang, L., Chen, Y., Nie, F., ... Zhang, G. (2015). Transcriptome profiling of the potato (*Solanum tuberosum* L.) plant under drought stress and water stimulus conditions. *PLoS One*, 10(5), e0128041.
- Gorji, A. M., Matyas, K. K., Duplecz, Z., Decsi, K., Cernak, I., Hoffmann, B., ... Polgar, Z. (2012). *In vitro* osmotic stress tolerance in potato and identification of major QTLs. *American Journal of Potato Research*, 89(6), 453–464.
- Goyer, A., Hamlin, L., Crosslin, J. M., Buchanan, A., & Chang, J. H. (2015). RNA-Seq analysis of resistant and susceptible potato varieties during the early stages of potato virus Y infection. *BMC Genomics*, 16(1), 472. Available from <https://doi.org/10.1186/s12864-015-1666-2>.
- Griffel, L. M., Delparte, D., & Edwards, J. (2018). Using Support vector machines classification to differentiate spectral signatures of potato plants infected with potato virus Y. *Computers and Electronics in Agriculture*, 153, 318–324.
- Habe, I., Miyatake, K., Nunome, T., Yamasaki, M., & Hayashi, T. (2019). QTL analysis of resistance to bacterial wilt caused by *Ralstonia solanacearum* in potato. *Breeding Science*, 69(4), 592–600.
- Hamilton, J. P., Hansey, C. N., & Whitty, B. R. (2011). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics*, 12, 302. Available from <https://doi.org/10.1186/1471-2164-12-302>.
- Hancock, R. D., Morris, W. L., Ducreux, L. J., Morris, J. A., Usman, M., Verrall, S. R., ... Hedley, P. E. (2014). Physiological, biochemical and molecular responses of the potato (*Solanum tuberosum* L.) plant to moderately elevated temperature. *Plant, Cell & Environment*, 37, 439–450.
- Hara-Skrzypiec, A., Śliwka, J., Jakuczun, H., & Zimnoch-Guzowska, E. (2018). QTL for tuber morphology traits in diploid potato. *Journal of Applied Genetics*, 59(2), 123–132.
- Hein, J. A., Sherrard, M. E., Manfredi, K. P., & Abebe, T. (2016). The fifth leaf and spike organs of barley (*Hordeum vulgare* L.) display different physiological and metabolic responses to drought stress. *BMC Plant Biology*, 16, 248. Available from <https://doi.org/10.1186/s12870-016-0922-1>.
- Hu, X., Hodén, K. P., Liao, Z., Dörfors, F., Åsman, A. K., & Dixelius, C. (2020). *Phytophthora infestans* Ago1-bound miRNA promotes potato late blight disease. *bioRxiv*. Available from <https://doi.org/10.1101/2020.01.28.924175>.
- Hu, Y. H., Ping, X. W., Xu, M. Z., Shan, W. X., & He, Y. (2016). Detection of late blight disease on potato leaves using hyperspectral imaging technique. *Spectroscopy and Spectral Analysis*, 36, 515–519.
- Jin, X., Zarco-Tejada, P., Schmidhalter, U., Reynolds, M. P., Hawkesford, M. J., Varshney, R. K., ... Li, S. (2021). High-throughput estimation of crop traits: A review of ground and aerial phenotyping platforms. *IEEE Geoscience and Remote Sensing Magazine*, 9, 200–231.
- Kaminski, K. P., Kørup, K., & Andersen, M. N. (2016). Next generation sequencing bulk segregant analysis of potato support that differential flux into the cholesterol and stigmaterol metabolite pools is important for steroidal glycoalkaloid content. *Potato Research*, 59, 81–97.
- Kikuchi, A., Huynh, H. D., Endo, T., & Kazuo, W. (2015). Review of recent transgenic studies on abiotic stress tolerance and future molecular breeding in potato. *Breeding Science*, 65(1), 85–102.
- Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L. S., & Paterson, A. H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science (Shannon, Ireland)*, 242, 14–22.
- Kitazumi, A., Kawahara, Y., Onda, T. S., De Koeber, D., & de los Reyes, B. G. (2015). Implications of miR166 and miR159 induction to the basal response mechanisms of an andigena potato (*Solanum tuberosum* subsp. andigena) to salinity stress, predicted from network models in *Arabidopsis*. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 58(1), 13–24.
- Klaassen, M. T., Willemsen, J. H., Vos, P. G., Visser, R. G. F., van Eck, H. J., Maliepaard, C. T., ... Luisa, M. (2019). Genome-wide association analysis in tetraploid potato reveals four QTLs for protein content. *Molecular Breeding*, 39(11), 151. Available from <https://doi.org/10.1007/s11032-019-1070-8>.
- Kloosterman, B., De Koeber, D., Griffiths, R., Flinn, B., Steuarnagel, B., Scholz, U., ... Bachem, C. W. B. (2008). Genes driving potato tuber initiation and growth: Identification based on transcriptional changes using the POCI array. *Functional & Integrative Genomics*, 8(4), 329–340.
- Kohl, M., Megger, D. A., Trippler, M., Meckel, H., Ahrens, M., & Bracht, T. (2014). A practical data processing workflow for multi-OMICS projects. *Biochimica et Biophysica Acta*, 1844, 52–62.
- Kyriakidou, M., Achakgari, S. R., Gálvez-López, J. H., Zhu, X., Tang, C. Y., Tai, H. H., ... David, S. M. V. (2020). Structural genome analysis in cultivated potato taxa. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 133(3), 951–966.

- Kyriakidou, M., Anglin, N. L., Ellis, D., Tai, H. H., & StrÅmvik, M. V. (2020). Genome assembly of six polyploid potato genomes. *Science Data*, 7 (1), 88. Available from <https://doi.org/10.1038/s41597-020-0428-4>.
- Lahner, B., Gong, J., Mahmoudian, M., Smith, E. L., & Abid, K. B. (2003). Genomic scale profiling of nutrient and trace elements in *Arabidopsis thaliana*. *Nature Biotechnology*, 2003(21), 1215–1221.
- Lakhotia, N., Joshi, G., & Bhardwaj, A. R. (2014). Identification and characterization of miRNAome in root, stem, leaf and tuber developmental stages of potato (*Solanum tuberosum* L.) by high-throughput sequencing. *BMC Plant Biology*, 14, 6. Available from <https://doi.org/10.1186/1471-2229-14-6>.
- Larsen, M. K., Guldstrand, J., Malene, M., Bennike, T. B., & Stensballe, A. (2016). Time-course investigation of *Phytophthora infestans* infection of potato leaf from three cultivars by quantitative proteomics. *Data in Brief*, 6, 238–248.
- Lehesranta, S. J., Davies, H. V., Shepherd, L. V. T., Koistinen, K. M., Massat, N., Nunan, N., ... Kärenlampi, S. O. (2006). Proteomic analysis of the potato tuber life cycle. *Proteomics*, 6, 6042–6052.
- Leisner, C. P., Hamilton, J. P., Crisovan, E., Manrique-Carpintero, N. C., & Marand, A. P. (2018). Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuberbearing potato species *Solanum chacoense*, reveals residual heterozygosity. *The Plant Journal*, 94, 562–570.
- Lekota, M., Modisane, K. J., Apostolides, Z., & van der Waals, J. E. (2020). Metabolomic fingerprinting of potato cultivars differing in susceptibility to *Spongopora subterranea* f. sp. subterranea root infection. *International Journal of Molecular Sciences*, 21(11), 3788. Available from <https://doi.org/10.3390/ijms21113788>.
- Li, J., Wang, Y., Wen, G., Li, G., Li, Z., Zhang, R., ... Xie, C. (2019). Mapping QTL underlying tuber starch content and plant maturity in tetraploid potato. *The Crop Journal*, 7(2), 261–272.
- Li, J., Zhu, L., Lu, G., Zhan, X. B., Lin, C. C., & Zheng, Z. Y. (2014). Curdian β -1,3-glucooligosaccharides induce the defense responses against *Phytophthora infestans* infection of potato (*Solanum tuberosum* L. cv. McCain G1) leaf cells. *PLoS One*, 9(5), e97197. Available from <https://doi.org/10.1371/journal.pone.0097197>.
- Li, L., Tacke, E., & Hofferbert, H. R. (2013). Validation of candidate gene markers for marker-assisted selection of potato cultivars with improved tuber quality. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 1039–1052.
- Li, L., Zou, Q., Deng, X., Peng, M. S., Huang, J., Lu, X. L., ... Wang, X. (2017). Comparative morphology, transcription, and proteomics study revealing the key molecular mechanism of camphor on the potato tuber sprouting effect. *International Journal of Molecular Sciences*, 18(11), 2280. Available from <https://doi.org/10.3390/ijms18112280>.
- Li, Q., Qin, Y., & Hu, X. (2020). Transcriptome analysis uncovers the gene expression profile of salt-stressed potato (*Solanum tuberosum* L.). *Scientific Reports*, 10, 5411. Available from <https://doi.org/10.1038/s41598-020-62057-0>.
- Liu, X., Chen, L., Shi, W., Xu, X., Li, Z., Liu, T., ... Song, B. (2020). Comparative transcriptome reveals distinct starch-sugar interconversion patterns in potato genotypes contrasting for cold-induced sweetening capacity. *Food Chemistry*, 127–150.
- Liu, Y., Lin-Wang, K., Deng, C., Warran, B., Wang, L., & Yu, B. (2015). Comparative transcriptome analysis of white and purple potato to identify genes involved in anthocyanin biosynthesis. *PLoS One*, 10(6), e0129148.
- Losifidis, M. (2011). *Evaluation of salt stress tolerance and ion content QTL analysis of potato grown in hydroponics and in the field*. Wageningen UR: Plant Breeding.
- Massa, A. N., Childs, K. L., Lin, H., Bryan, G. J., Giuliano, G., & Buell, C. R. (2011). The transcriptome of the reference potato genome *Solanum tuberosum* Group Phureja clone DM1–3 516R44. *PLoS One*, 6(10), e26801.
- Mathieu, D., Morin, P. J., Lyons, N. C., Sébastien, B., & Pier, J. M. (2017). Identification of differentially expressed miRNAs in Colorado potato beetles (*Leptinotarsa decemlineata* (Say)) exposed to imidacloprid. *International Journal of Molecular Sciences*, 18, 2728. Available from <https://doi.org/10.3390/ijms18122728>.
- Michelmore, R., Reyes Chin-Wo, S., & Kozik, A. (2016). Improvement of the genome assembly of lettuce (*Lactuca sativa*) using dovetail/in vitro proximity ligation. In *Plant and animal genome XXIV conference*.
- Milella, A., Marani, R., Petitti, A., & Reina, G. (2019). In-field high throughput grapevine phenotyping with a consumer-grade depth camera. *Computers and Electronics in Agriculture*, 156, 293–306.
- Minhas, J. S. (2012). *Potato: Production strategies under abiotic stress in improving crop resistance to abiotic stress* (1st ed.). Weinheim: Wiley, [Chapter 45].
- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419. Available from <https://doi.org/10.3389/fpls.2016.01419>.
- Moon, K. B., Ahn, D. J., Park, J. S., Jung, W. Y., Cho, H. S., Kim, H. R., ... Kim, H. S. (2018). Transcriptome profiling and characterization of drought-tolerant potato plant (*Solanum tuberosum* L.). *Molecules and Cells*, 41(11), 979–992.
- Morozova, O., Hirst, M., & Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics*, 10, 135–151.
- Nie, X., Sutherland, D., & Dickison, V. (2016). Development and validation of high resolution melting markers derived from *Ry_{sto}* STS markers for high throughput marker-assisted selection of potato carrying *Ry_{sto}*. *Phytopathology*, 106, 1366–1375.
- Oppenheim, D., Shani, G., Erlich, O., & Tsror, L. (2019). Using deep learning for image-based potato tuber disease detection. *Phytopathology*, 109, 1083–1087.
- Osmani, Z., Sabet, M. S., Shams-Bakhsh, M., Moieni, A., Vahabi, K., & Wehling, P. (2019). Virus-specific and common transcriptomic responses of potato against PVY, PVA and PLRV using microarray meta-analysis. *Plant Breeding*, 138, 216–228.

- Outten, C. E., & O'Halloran, T. V. (2001). Femtomolar sensitivity of metalloregulatory proteins controlling zinc homeostasis. *Science (New York, N. Y.)*, 292, 2488–2492.
- Plischke, A., Choi, Y. H., Brakefield, P. M., Klinkhamer, P. G. L., & Bruinsma, M. (2012). Metabolomic plasticity in gm and non-GM potato leaves in response to aphid herbivory and virus infection. *Journal of Agricultural and Food Chemistry*, 60(6), 488–1493.
- Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome*, 92. Available from <https://doi.org/10.19080/AIBM.2019.14.555891>.
- Potato Genome Sequencing Consortium. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475, 189–195.
- Puzanskiy, R. K., Yemelyanov, V. V., Gavrilenko, T. A., & Shishova, M. F. (2017). The perspectives of metabolomic studies of potato plants. *Russian Journal of Genetics: Applied Research*, 7(7), 744–756.
- Rajamaki, M. L., Sikorskaite-Gudziuniene, S., Sarmah, N., Varjosalo, M., & Valkonen, J. P. T. (2020). Nuclear proteome of virus-infected and healthy potato leaves. *BMC Plant Biology*, 20(1), 355. Available from <https://doi.org/10.1186/s12870-020-02561-7>.
- Ramakrishnan, A. P., Ritland, C. E., & Blas Sevillano, R. H. (2015). Review of potato molecular markers to enhance trait selection. *American Journal of Potato Research*, 92, 455–472.
- Ray, S. S., Jain, N., Arora, R. K., Chavan, S., & Panigrahy, S. (2011). Utility of hyperspectral data for potato late blight disease detection. *Journal of the Indian Society of Remote Sensing*, 39, 161–169.
- Regan, S., Gustafson, V., Rothwell, C., Sardana, R., Flinn, B., Mallubhotla, S., . . . De Koeper, D. (2006). Finding the perfect potato: Using functional genomics to improve disease resistance and tuber quality traits. *Canadian Journal of Plant Pathology*, 28, S247–S255.
- Rensink, W. A., Iobst, S., Hart, A., Stegalkina, S., Liu, J., & Robin, B. C. (2005a). Gene expression profiling of potato responses to cold, heat, and salt stress. *Functional & Integrative Genomics*, 5(4), 201–207.
- Rensink, W. A., Iobst, S., Hart, A., Stegalkina, S., Liu, J., & Robin, B. C. (2005b). Analyzing the potato abiotic stress transcriptome using expressed sequence tags. *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 48(4), 598–605.
- Ritter, E., Debener, T., & Barone, A. (1991). RFLP mapping on potato chromosomes of two genes controlling extreme resistance to potato virus X (PVX). *Molecular & General Genetics: MGG*, 227, 81–85.
- Rodziewicz, P., Swarczewicz, B., Chmielewska, K., Wojakowska, A., & Stobiecki, M. (2014). Influence of abiotic stresses on plant proteome and metabolome changes. *Acta Physiologiae Plantarum/Polish Academy of Sciences, Committee of Plant Physiology Genetics and Breeding*, 1–19.
- Salvato, F. J. F., Havelund, M., Chen, R. S. P., Rao, A., Rogowska-Wrzęsinska, O. N., Jensen, D. R., . . . Thelen, I. M. (2014). The potato tuber mitochondrial proteome. *Plant Physiology*, 164, 637–653.
- Samuelsson, L. M., & Larsson, D. G. (2008). Contributions from metabolomics to fish research. *Molecular Biosystems*, 4(10), 974. Available from <https://doi.org/10.1039/b804196b>.
- Santa, J. D., Berdugo-Cely, J., Cely-Pardo, L., Soto-Suárez, M., Mosquera, T., & Galeano, M. C. H. (2018). QTL analysis reveals quantitative resistant loci for *Phytophthora infestans* and *Teciasolanivora* in tetraploid potato (*Solanum tuberosum* L.). *PLoS One*, 13(7), e0199716.
- Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., & Selbig, J. (2014). Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics (Oxford, England)*, 20, 2447–2454.
- Schönhals, E. M., Ortega, F., & Barandalla, L. (2016). Identification and reproducibility of diagnostic DNA markers for tuber starch and yield optimization in a novel association mapping population of potato (*Solanum tuberosum* L.). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 129, 767–785.
- Schumacher, A., Rujan, T., & Hoefkens, J. (2014). A collaborative approach to develop a multi-omics data analytics platform for translational research. *Applied & Translational Genomics*, 3, 105–108.
- Sha, Z., Oka, N., Watanabe, T., Tampubolon, B. D., Okazaki, K., Osaki, M., . . . Shinano, T. (2012). Ionome of soybean seed affected by previous cropping with mycorrhizal plant and manure application. *Journal of Agricultural and Food Chemistry*, 60, 9543–9552.
- Simko, I., Haynes, K. G., & Ewing, E. E. (2004). Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. *Molecular Genetics and Genomics: MGG*, 271, 522–531.
- Sołtys-Kalina, D., Szajko, K., Wasilewicz-Flis, I., Mańkowski, D., Marczewski, W., & Śliwka, J. (2020). Quantitative trait loci for starch-corrected chip color after harvest, cold storage and after reconditioning mapped in diploid potato. *Molecular Genetics and Genomics: MGG*, 295(1), 209–219.
- Song, Y. S., Hepting, L., & Schweizer, G. (2005). Mapping of extreme resistance to PVY (Rysto) on chromosome XII using anther-culture-derived primary dihaploid potato lines. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 111, 879–887.
- Spöner, D. M., Ghislain, M., Reinhard, S., Jansky, S. H., & Gavrilenko, T. (2014). Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. *The Botanical Review*, 80(4), 283–383.
- Sprenger, H., Kurowsky, C., Horn, R., Erban, A., Seddig, S., Rudack, K., . . . Kopka, J. (2016). The drought response of potato reference cultivars with contrasting tolerance. *Plant, Cell & Environment*, 39(11), 2370–2389.
- Suzuki, N., Rivero, R. M., Shulaev, V., Blumwald, E., & Mittler, R. (2014). Abiotic and biotic stress combinations. *The New Phytologist*, 203, 32–43.
- Szajko, K., Yin, Z., & Marczewski, W. (2019). Accumulation of miRNA and mRNA targets in potato leaves displaying temperature-dependent responses to potato virus Y. *Potato Research*, 62, 379–392.
- Szpunar, J. (2004). Metallomics: A new frontier in analytical chemistry. *Analytical and Bioanalytical Chemistry*, 378, 54–56.
- Tai, H. H., Lagüe, M., Thomson, S., Arousseau, F., Neilson, J., Murphy, A., . . . Jacobs, J. M. E. (2020). Tuber transcriptome profiling of eight potato cultivars with different cold-induced sweetening responses to cold storage. *Plant Physiology and Biochemistry: PPB/Societe Francaise de Physiologie Vegetale*, 146, 163–176.

- Tang, R., Gupta, S. K., Niu, S., Li, X. Q., Yang, Q., Chen, G., ... Haroon, M. (2020). Transcriptome analysis of heat stress response genes in potato leaves. *Molecular Biology Reports*, *47*(6), 4311–4321.
- Tiwari, J. K., Buckseth, T., Devi, S., Varshney, S., Sahu, S., Patil, V. U., ... Rai, A. (2020). Physiological and genome-wide RNA-sequencing analyses identify candidate genes in a nitrogen-use efficient potato cv. Kufri Gaurav. *Plant Physiology and Biochemistry: PPB/Societe Francaise de Physiologie Vegetale*, *154*, 171–183.
- Tiwari, J. K., Buckseth, T., Singh, R. K., Kumar, M., & Kant, S. (2020). Prospects of improving nitrogen use efficiency in potato: Lessons from transgenics to genome editing strategies in plants. *Frontiers in Plant Science*, *11*, 597481. Available from <https://doi.org/10.3389/fpls.2020.597481>.
- Tiwari, J. K., Buckseth, T., Zinta, R., Saraswati, A., Singh, R. K., Rawat, S., ... Chakrabarti, S. K. (2020a). Genome-wide identification and characterization of microRNAs by small RNA sequencing for low nitrogen stress in potato. *PLoS One*, *15*(5), e0233076.
- Tiwari, J. K., Buckseth, T., Zinta, R., Saraswati, A., Singh, R. K., Rawat, S., ... Chakrabarti, S. K. (2020b). Transcriptome analysis of potato shoots, roots and stolons under nitrogen stress. *Scientific Reports*, *10*, 1152. Available from <https://doi.org/10.1038/s41598-020-58167-4>.
- Tiwari, J. K., Challam, C., Chakrabarti, S. K., & Feingold, S. E. (2020). Climate smart potato: An integrated breeding, genomics and phenomics approach. In C. Kole (Ed.), *Genomic designing of climate smart vegetable crops* (pp. 1–46). Switzerland: Springer Nature.
- Tiwari, J. K., Devi, S., Buckseth, T., Ali, N., Singh, R. K., Zinta, R., ... Chakrabarti, S. K. (2020). Precision phenotyping of contrasting potato (*Solanum tuberosum* L.) varieties in a novel aeroponics system for improving nitrogen use efficiency: In search of key traits and genes. *Journal of Integrative Agriculture*, *19*(1), 51–61.
- Tiwari, J. K., Plett, D., Garnett, T., Chakrabarti, S. K., & Singh, R. K. (2018). Integrated genomics, physiology and breeding approaches for improving nitrogen use efficiency in potato: Translating knowledge from other crops. *Functional Plant Biology: FPB*, *45*, 587–605.
- Tiwari, J. K., Siddappa, S., & Singh, B. P. (2013). Molecular markers for late blight resistance breeding of potato: An update. *Plant Breeding*, *132*, 237–245.
- Toubiana, D., Fernie, A. R., Nikoloski, Z., & Fait, A. (2013). Network analysis: Tackling complex data to study plant metabolism. *Trends in Biotechnology*, *31*, 29–36.
- Uitdewilligen, J., Wolters, A. M. A., & D'hoop, B. B. (2013). A Next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One*, *8*, 10–14.
- Uribe, P., Jansky, S., & Halterman, D. (2014). Two CAPS markers predict Verticillium wilt resistance in wild *Solanum* species. *Molecular Breeding*, *33*, 465–476.
- Valin, H. S., Ronald, D., Mensbrugge, V. D., Nelson, D., Gerald, C. A., Blanc, E., ... Willenbockel, D. (2014). The future of food demand: Understanding differences in global economic models. *Agricultural Economics*, *45*(1), 51–67.
- Vos, P. G., Uitdewilligen, J., & Voorrips, R. E. (2015). Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): An insight into the breeding history. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *128*, 2387–2401.
- Wang-Pruski, G., & Schofield, A. (2012). *Potato: Improving crop productivity and abiotic stress tolerance in improving crop resistance to Abiotic Stress* (1st ed.). Weinheim: Wiley, [chapter 44].
- Williams, R. J. P. (2001). Chemical selection of elements by cells. *Coordination Chemistry Reviews*, 216–217.
- Xiao, C., Gao, J., Zhang, Y., Wang, Z., Zhang, D., Chen, Q., ... Shen, Y. (2019). Quantitative proteomics of potato leaves infected with *Phytophthora infestans* provides insights into coordinated and altered protein expression during early and late disease stages. *International Journal of Molecular Sciences*, *136*. Available from <https://doi.org/10.3390/ijms20010136>.
- Yang, X., Guo, X., Yang, Y., Ye, P., Xiong, X., Liu, J., ... Li, G. (2018). Gene profiling in late blight resistance in potato genotype SD20. *International Journal of Molecular Sciences*, *19*(6), 1728. Available from <https://doi.org/10.3390/ijms19061728>.
- Yang, Y., Qiang, X., Owsiany, K., Zhang, S., Thannhauser, T. W., & Li, L. (2011). Evaluation of different multidimensional LC–MS/MS pipelines for isobaric tags for relative and absolute quantitation (iTRAQ)-based proteomic analysis of potato tubers in response to cold storage. *Journal of Proteome Research*, *10*, 4647–4660.
- Yin, C. C., Ma, B., Collinge, D. P., Pogson, B. J., He, S. J., & Xiong, Q. (2015). Ethylene responses in rice roots and coleoptiles are differentially regulated by a carotenoid isomerase-mediated abscisic acid pathway. *The Plant Cell*, *27*, 1061–1081.
- Yu, W., Choi, J. S., Upadhyaya, C. P., Kwon, S. O., Gururani, M. A., Nookaraju, A., ... Ajappala, H. (2012). Dynamic proteomic profile of potato tuber during its in vitro development. *Plant Science (Shannon, Ireland)*, *195*, 1–9.
- Zhang, N., Yang, J., Wang, Z., Wen, Y., Wang, J., He, W., ... Wang, D. (2014). Identification of novel and conserved microRNAs related to drought stress in potato by deep sequencing. *PLoS One*, *9*(4), e95489.
- Zhang, W., Luo, Y., Gong, X., Zeng, W., & Li, S. (2009). Computational identification of 48 potato microRNAs and their targets. *Computational Biology and Chemistry*, *33*, 84–93.
- Zhang, Y. T., Zhou, D. Q., Su, Y., Yu, P., Zhou, X. G., & Yao, C. X. (2013). Proteome analysis of potato drought resistance variety in Ninglang 182 leaves under drought stress. *Hereditas*, *35*(5), 666–672.
- Zhao, J., Liu, Q., Hu, P., Jia, Q., Liu, N., Yin, K., ... Liu, Y. (2016). An efficient Potato virus X -based microRNA silencing in *Nicotiana benthamiana*. *Scientific Reports*, *6*, 20573. Available from <https://doi.org/10.1016/j.virusres.2012.02.012>.
- Zou, C., Wang, P., & Xu, Y. (2016). Bulk sample analysis in genetics, genomics and crop improvement. *Plant Biotechnology Journal*, *14*, 1941–1955.
- Zuluaga, A. P., Solé, M., Lu, H., Góngora-Castillo, E., Vaillancourt, B., Coll, N., ... Valls, M. (2015). Transcriptome responses to *Ralstonia solanacearum* infection in the roots of the wild potato *Solanum commersonii*. *BMC Genomics*, *16*(1), 246. Available from <https://doi.org/10.1186/s12864-015-1460-1>.

This page intentionally left blank

Tea plant genome sequencing: prospect for crop improvement using genomics tools

Pradosh Mahadani¹ and Basant K. Tiwary²

¹National Tea Research Foundation, Tea Board, Kolkata, India, ²Department of Bioinformatics, Pondicherry University, Pondicherry, India

22.1 Introduction

The tea plant is an important woody cash crop with three races, *Camellia sinensis*, *Camellia assamica*, and *Cambodiensis*. Tea is the most popular consumed beverage due to its medicinal, refreshing, and mild stimulant effects and is believed to originate in China and Southeast Asia. The tea manufacturing process has categorized tea into four types of basic types, “unfermented green tea,” “fully fermented black tea,” “partially fermented oolong tea” (all three are made from apical shoots, i.e., two leaves and a bud), and “white tea from tea leaf buds only” (Fig. 22.1). Tea is manufactured through a sequential manufacturing process in five stages, withering, rolling, fermenting, drying, and sorting as taken out from the plant’s green leaves. The worldwide production of tea is about 6150 million kg in 2019 and India is the second-largest tea producer of tea after China. Tea occupies an important place among cash crops due to its major contribution to the country’s economy and it further provides livelihood to a large number of people. In 1823 Robert Bruce discovered wild tea plants in the upper Brahmaputra Valley, India. More than 120 varieties of tea are cultivated in India, such as Assam, West Bengal, Sikkim, Kerala, Tamil Nadu, Himachal Pradesh, and Tripura (Board, 2019). Three primary secondary metabolites (flavonoids, theanine, and caffeine) are responsible for astringent, umami, and bitter taste of tea and the overall tea quality. The abundant modern therapeutic research in this respect and the evidences supporting tea drinking’s health benefits forms the scientific basis for this belief and claim. Among the flavonoids, catechin is major derivative, which is mixture of catechins, gallocatechins, epicatechins, and epigallocatechins (Higdon & Frei, 2003). Both the catechins and theaflavins, two significant tea ingredients, are antioxidants. Recent research suggested the health benefits of these polyphenols include protection against viruses, cardiovascular ailments, and decreased risk of several types of cancers, skin damage, long-term diabetes management, weight management, in addition to other benefits (Chacko et al., 2010).

Although tea is an important cash crop, it has attained genomic saturation through repeated conventional breeding. High heterozygosity, genetic erosion, long gestation period, lack of mutants, and nonavailability of the structured population are primary bottlenecks in tea breeding (Mukhopadhyay, Mondal, & Chand, 2016). In this context, genomics-assisted breeding (GAB) is the most appropriate breeding choice for enhancing the tea production. The availability of genomic resources, genetic linkage maps, and mapped and cloned gene information are essential to perform GAB in any crop (Fig. 22.2). Over the last decade, next-generation sequencing has become an indispensable tool to generate large-scale genomic resources for crop improvement. Rapid progress in sequencing techniques, bioinformatics pipelines, software, and strategies has allowed sequencing the whole genome and transcriptome at a very cheaper cost.

Since 2010, more than 2700 Short Read Archive (SRA) files from 177 bioprojects have been generated from tea plants’ tissues (Fig. 22.3). Due to the explosion of tea genomic resources during the last 5 years, few novel specialized databases have been developed in China and India. More databases are likely to operational shortly. Tea Plant Information Archive (TPIA; <http://tpia.teaplant.org>) is the first integrative and specially designed web-accessible database on tea plant. TPIA hosts annotated tea plant genome, well-organized transcriptomes, gene



FIGURE 22.1 Two leaves and buds from tea plants used in manufacturing of quality tea.

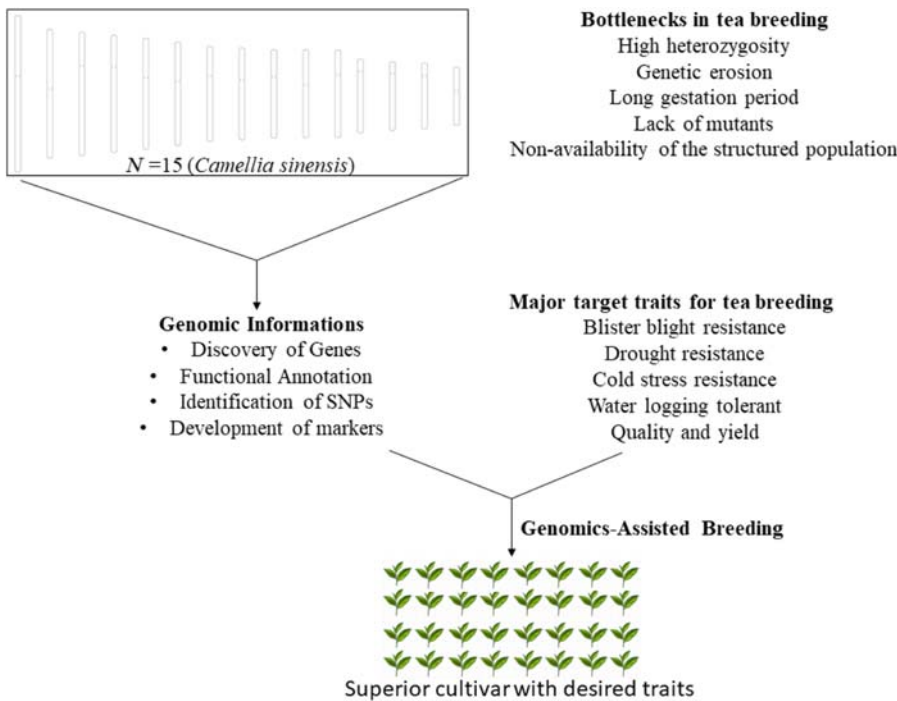


FIGURE 22.2 Schematic diagram for tea improvement program with recent advancements of genomic data.

expressions pattern, orthologs and characteristic metabolites of tea quality, massive transcription factors (TF), polymorphic simple sequence repeats, single-nucleotide polymorphism (SNP), etc. Anhui Agricultural University, China, developed this knowledgebase and served as a central gateway for tea research community (Xia, E.H. et al., 2019). Besides that, TeaMiD, a specialized database for Simple Sequence Repeat (SSR) markers for tea is developed by the ICAR-National Institute of Plant Biotechnology, New Delhi. They identified 935,547 SSRs from *C. sinensis* var *sinensis* (CSS) genome and other publicly available genomic resources (Dubey et al., 2020). A gene coexpression network for tea plant-based database, TeaCoN is recently developed by Zhang, R. et al. (2020). It will help in understanding the functional role of candidate genes in tea plants. This chapter is focused on the different applications of next-generation sequencing technology in tea research and crop improvement program. A comparative account of available tea genomes is discussed with their merits and demerits as a reference genome.

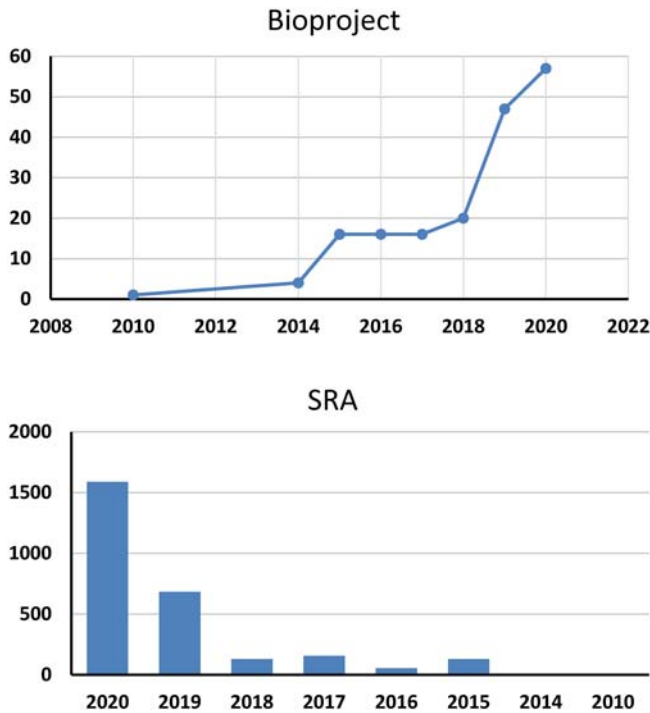


FIGURE 22.3 Number of bioprojects and short read archive submitted during the last 10 years (since 2010–20).

22.2 Whole-genome sequencing of tea plant

In recent years the whole-genome sequencing of two tea plant varieties, CSS and *C. sinensis* var *assamica* (CSA), has been attempted by various research groups with different purposes and strategies. The whole-genome sequencing of an organism is more accessible and less time-consuming with the advancement of the next-generation of sequencing (NGS) technologies. However, the genome assembly of *C. sinensis* ($n = 15$) is a complicated and tedious process due to large-size genome (~ 3 GB), highly repetitive DNA content, and high level of heterozygosity. Sequencing the whole genome of tea plant was achieved using different platforms, library preparation, and assembly techniques (Table 22.1). Short sequence read assembly; long sequence read assembly, hybrid assembly, Hi-C, etc. are applied to decipher the 15 pseudomolecules or chromosome-scale genome (Fig. 22.4). Xia et al. (2017) first decoded the draft genome of *C. sinensis* var *assamica* (Xia et al., 2017). By the end of 2020 a total of eight whole genomes of tea were available in the public domain (Table 22.1). Some of these projects have described a well-annotated genome and chromosome-scale tea genome necessary for breeding, evolution, and adaptation of tea plant. All the genome assemblies have about ~ 3 Gb with an average $\sim 200\times$ depth and scaffold N50 sizes between 449 kb and 218 mb. There are 32,311–53,512 predicted protein-coding genes in the tea plant genomes.

A high-quality reference genome or chromosome-scale genome assembly is a challenging and essential requirement for identifying quantitative trait loci (QTLs) useful in breeding programs. Hi-C technology is a powerful technique that has been developed to guide genome assembly. Chen, J.D. et al. (2020) has recently published a chromosome-scale genome of tea plant to identify the useful genes or QTLs involved in secondary metabolites and understand the evolutionary role of duplicated genes in diversifying tea plants. They also resequenced 139 tea accessions worldwide among tea-growing countries except for India to understand the origin and evolution of tea plants (Chen, J.D. et al., 2020). Besides, Xia et al. (2020) and Wang, X. et al. (2020) have resequenced 81 and 139 accessions of tea plants with a focus on the high-resolution genomic variations (62,52,201 and 21,88,70,000 SNPs, respectively) that might play a vital role in the tea breeding research, especially in the marker-assisted breeding program and overall genetic improvement of the crop (Wang, X. et al., 2020; Xia et al., 2020). Moreover, Zhang, W. et al. (2020) reported a chromosome-scale genome from an ancient tea species known as DASZ, focusing on evolution of this ancient species (Zhang, W. et al., 2020).

TABLE 22.1 Comparative account of available tea genomes (assembled up to September, 2020).

	Xia et al. (2017)	Wei et al. (2018)	Xia, E. et al. (2019)	Mondal et al. (2019)	Xia et al. (2020)	Zhang, Q.J. et al. (2020)	Chen, J.D. et al. (2020)
Tea cultivar name	CSA (Yunkang 10)	CSS (Shuchazao)	CSS (Shuchazao)	<i>C. assamica</i> (TV1)	CSS (Shuchazao)	CSS (<i>Biyun</i>)	CSS (Shuchazao)
Level of heterozygosity	–	Low (2.7%) RAD-Seq	Low (2.7%) RAD-Seq	–	–	Low (~1.22%)	–
Platform used	HiSeq 2000 (Illumina)	HiSeq 2500 PacBio RSII	HiSeq 2500 PacBio RSII	HiSeq 2500 PacBio RSII	PacBio sequel and HiSeq X 10	PacBio RSII and HiSeq X 10	HiSeq 4000
Library used	Paired-end metapair	Paired-end metapair, 10, 20 kb	Paired-end metapair, 10, 20 kb	Paired-end metapair, 10, 20 kb	Paired-end, 20 kb, Hi-C	Paired, Hi-C	Hi-C
Estimated genome size (Gb)	3	3.14	3.08	3	3	3.25	3.2
Name of genome assemblers	Platanus, SSPACE	SOAPdenovo	SOAPdenovo, Platanus, hybrid approach	Platanus, GapCloser, Sealer, LACHESIS	FALCON, Pilon, LACHESIS	FALCON, Purge Haplotigs, SSPACE, LACHESIS, JUICERBOX	Juicer pipeline, 3D-DNA
Amount of raw data (Gb)	~ 707.88	1450.4	2249.16	–	–	417.95	337.8
Depth (X)	159.43	436	464.21	149.29	87.2	127.66	113
No scaffolds	37,618	14,051	14,051	14,824	?	4153	14,412
	Xia et al. (2017)	Wei et al. (2018)	Xia, E. et al. (2019)	Mondal et al. (2019)	Xia et al. (2020)	Zhang, Q.J. et al. (2020)	Chen, J.D. et al. (2020)
N50 of scaffolds (bp)	4,49,457	13,97,810	13,97,810	5,38,958	–	19,56,80,000	21,81,15,851
Assembled genome size (Gb)	~3.02	~3.1	2.98	2.93	2.94	~2.92	2.98
Coverage of assembled genome (%)	~98%	93%	95.07%	97.66%	98%	89.85	94.07
GC %	42.31	–	37.84	39.57	–	38.24	–
No. of predicted protein-coding gene	–	33,932	53,512	–	50525	40,812	32,311
Repeat sequences (%)	80.89	64	64.42	71.87	86.87	74.13	–
References	Xia et al. (2017)	Wei et al. (2018)	Xia, E. et al. (2019)	Mondal et al. (2019)	Xia et al. (2020)	Zhang, Q.J. et al. (2020)	Chen, J.D. et al. (2020)
Data repository	Bioproject: PRJNA381277	pcsb.ahau.edu.cn.8080	TPIA	Bioproject: PRJNA597714	Bioproject: PRJCA002071	Bioproject: PRJNA596054	Bioproject: PRJNA646044

CSA, *Camellia sinensis* var *assamica*; CSS, *Camellia sinensis* var *sinensis*; TPIA, Tea Plant Information Archive.

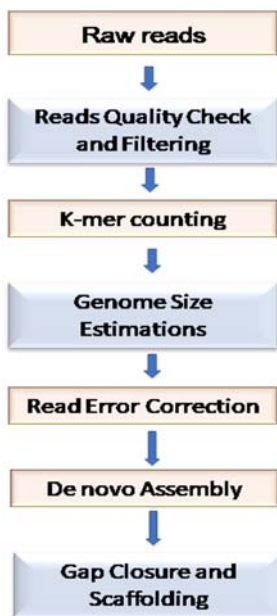


FIGURE 22.4 Various steps in the genome assembly process.

22.3 Identification and characterization of gene families

The first whole-genome sequence of tea plant was published in 2017 and the resulted post-tea genome era has allowed an excellent opportunity for researchers to systematically analyze and characterize the gene families in the tea genome. Recent studies reported important gene families in the tea genome and have significantly improved our understanding of their biological functions. For example, amino acid permeases (AAPs) play a crucial role in the uptake, transport, and distribution of amino acids and are well-known amino acid transporters. A total of 9 AAP genes are identified from the tea plant's genome and classified into three distinct groups based on their structures and conserved motifs. The expression pattern of CsAAPs is specific to different plant tissues, with five genes showing higher expression levels in the stem, six genes exhibiting higher expression in leaves, and the remaining genes showed higher expression in the root. Interestingly, the CsAAP-19 gene is exclusively expressed in the root (Duan et al., 2020). Genome-wide characteristics of different gene families and their descriptions are given in Table 22.2.

22.4 Tea transcriptome sequencing

RNA-Seq is a next-generation sequencing tool to understand the transcriptome of an organism and to decipher genomic functions, differential gene expressions, and posttranslational molecular mechanisms (Fig. 22.5) (Garg & Jain, 2013; Wang, Gerstein, & Snyder, 2009). The RNA-Seq methodology includes RNA extraction, preparation of sequencing library by cDNA synthesis followed by platform-specific adapter ligation. The libraries are usually sequenced as per the required read depth. Short read sequencing usually generates 200–500 bp long and an average of 20–30 million reads per sample. Iso-Seq technology from PacBio has the ability to generate full-length transcripts, though are also more error-prone. The overall bioinformatics analysis includes sequence quality check, trimming, aligning, and assembly of transcripts, quantifying the reads and studying the changes in the expression of genes across samples. Shi et al. (2011) first reported major metabolic pathways in tea plants by analyzing the high-throughput Illumina RNA-Seq (Shi et al., 2011). About 150 bioprojects have been submitted in the NCBI-SRA database allowing transcriptome analysis of tea for gene discovery and posttranscriptional molecular mechanisms in the tea genome.

Most of the transcriptome studies focused on discovering the functional genes and their involvement in the regulatory pathways responding to biotic and abiotic stresses affecting tea quality. The overall growth of tea plants and the quality of leaves are adversely affected due to biotic stress factors and other abiotic factors such as low temperature, heat, and drought. The secondary metabolites like flavonoids, caffeine, and theanine are important components determining tea quality. Guo et al. (2017) identified key genes involved in catechins biosynthesis using transcriptome analysis. The comparative transcriptional and metabolite profiles revealed that *PAL*, *C4H*, *F3H*, *LAR*, and *ANS* are critical genes for catechin biosynthesis during different leaf development stages (Guo et al., 2017).

TABLE 22.2 List of important gene families identified in tea genome.

Gene family	No. of genes	Function	References
Amino acid permease	19	Transportation of amino acids	Duan et al. (2020)
Serine carboxypeptidase-like acyltransferase (SCPL)	47	Encoded galloylated catechins	Ahmad et al. (2020)
C-repeat binding factor	6	Encoded transcriptional activators and role on cold tolerance	Hu et al. (2020)
Polyamine oxidase	7	Growth and development under environmental stress.	Li, M. et al. (2020)
WRKY	56	Diverse regulation and multiple stress responses	Shen et al., 2020; Wang et al. (2019)
Heat stress factors (Hsf)	25	In signal transduction pathways operating in response to environmental stresses	Zhang et al. (2020a)
SBP-box transcription factors	25	Transcription factor plays important role in the process of resisting abiotic stress.	Zhang, D. et al. (2020)
Metal-tolerance proteins (MTP)	13	MTPs are mainly involved in transporting Mn, Zn, and Fe	Zhang et al. (2020b)
DNA-binding one zinc finger (Dof)	16	Transcription factors are important for seed development, hormone regulation, and defense against abiotic stress	Yu et al. (2020)
Voltage-gated chloride channel	8	Transporting NO ³⁻ , Cl, and other monovalent anions	Xing et al. (2020)
Mitogen-activated protein kinase	21	Fundamental pathway in organisms for signal transduction	Chatterjee et al. (2020)
SABATH methyltransferases	32	Convert plant small-molecule metabolites into volatile methyl ester's tea plant defense responses	Guo et al. (2020)
PYL-PP2C-SnRK2s	106	PYL-PP2C-SnRK2s were associated with changes of leaf color and the response of <i>Camellia sinensis</i> to drought and salt stressors	Xu et al. (2020)
Cytosine-5 DNA methyltransferase and DNA demethylase	8 + 4	Abiotic stress and the potential functions of these two gene families in affecting tea flavor during tea withering processing	Wang, Y. et al. (2020)

Further, a coexpression analysis revealed 30 TF involved in the regulation of catechin biosynthesis. Li et al. (2015) studied the gene regulation involved in the secondary metabolite biosynthetic pathways using RNA-Seq. Tissue-specific expressed genes are identified from 13 different tissues, including leaves and buds at various developmental stages and tissue samples of stems, roots, flowers, and seeds. The study characterized the expression patterns of 206 unigenes involved in the flavonoid, caffeine, and theanine pathway. A total of 67 TFs are related to flavonoid, caffeine biosynthesis pathway and 22 TFs associated with the flavonoid, caffeine biosynthesis pathway. Moreover, one TF (c113397.0.1) from NAC TF is related to all three pathways.

Several studies provided a deep understanding of molecular mechanism involved in cold adaptation using transcription profile. *C. sinensis* var *sinensis* (CSS) showed high cold resistance after acclimation than the CSA. However, the winter dormancy and banji dormancy in tea plants are manifestations of suspension of development under unfavorable conditions. The tea plant undergoes a dormancy period when apical bud growth almost ceases, eventually leading to very low commercial yield. Low temperatures usually prevail during the winter dormancy of tea plants. The tea plants are expected to be under a complete dormancy period when the winter day is shorter than a critical day-length of about 11 hours 15 minutes continuing for at least 6 weeks. The dormancy period is directly proportional to the length of short days. These studies have provided a global transcriptome profile of winter dormancy and related regulatory mechanisms in the tea plant (Hao et al., 2017).

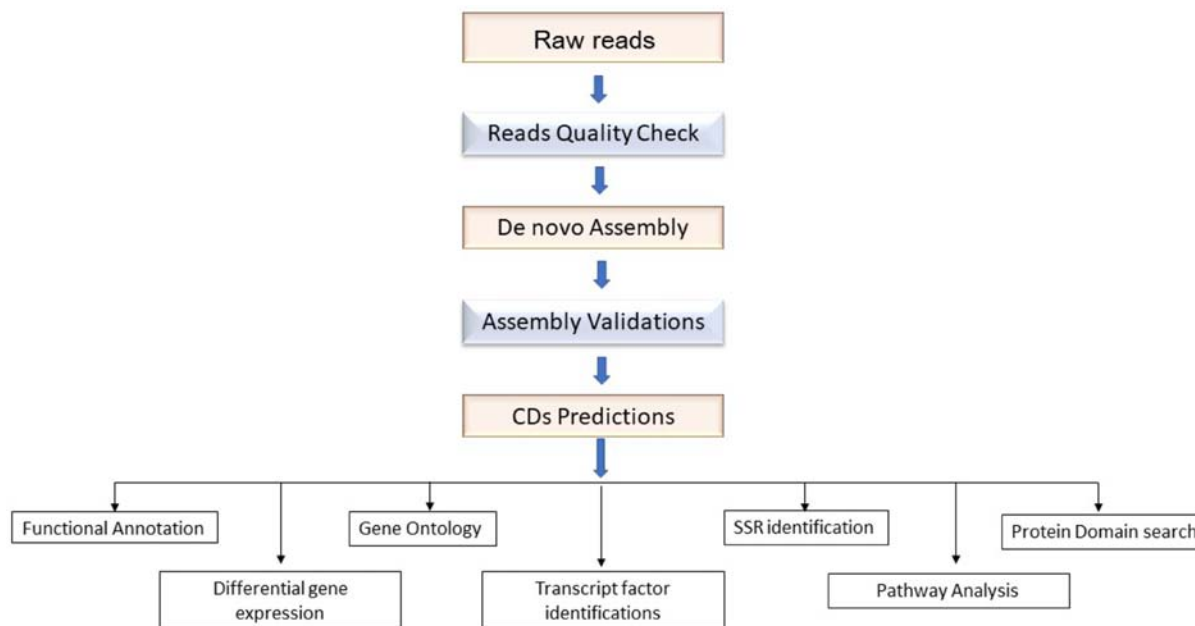


FIGURE 22.5 Basic steps of RNA-Seq analysis to achieve research goals.

TABLE 22.3 Selected transcriptome profile studies on tea and major pathways involve in stress management of tea.

Area of research	Major pathways/genes involved	References
Defense role of the unfested adjacent leaf by tea geometrids	JA (jasmonic acid), SA (salicylic acid), and ET (ethylene) synthesis pathway	Zhou et al. (2020)
Cold adaptation	Photosynthesis, plant hormonal signal transduction, and transcriptional regulation of plant–pathogen interaction	Li et al. (2019)
Sucrose treatment	Glutathione S-transferase, ATP-binding cassette transporters (ABC transporter), and MATE transporter	Qiao et al. (2019)
Selenium accumulation	Ribosome and protein processing in endoplasmic reticulum, sulfur metabolism, glutathione metabolism, selenocompound metabolism, and plant hormone signal transduction	Cao et al. (2018)
Bud dormancy	Epigenetic mechanism, phytohormone signaling, callose-related cellular communication regulation	Hao et al. (2017)
Drought and salinity stress	Starch and sucrose metabolism, plant hormonal signal transduction, photosynthesis, photosynthesis-antenna proteins, galactose metabolism, etc.	Zhang et al. (2017)
Nitrogen utilization	AMT, NRT, and AQP genes are involved in N uptake. GOGAT and GS genes are involved in N assimilation	Li, W. et al., (2017)
Aluminum tolerance	Transporters, transcription factors, cytochrome P450, ubiquitin ligase, organic acid biosynthesis, heat shock proteins	Li, Y. et al. (2017)
Seasonal variations	Catechin biosynthesis, caffeine biosynthesis/catabolism, phytohormones, histone, and DNA modification	Kumar et al. (2016)
Self-incompatibility	Plant hormone signal transduction, plant–pathogen interaction, flavonoid biosynthesis, calcium signaling pathway, and ubiquitin-mediated proteolysis	Zhang et al. (2016)
Blister blight defenses	R genes, defense-related enzymes, retrotransposons, transcription factors, and other defense-associated molecules	Jayaswall et al. (2016)
Methyl jasmonate-treated	α -Lenolenic acid degradation, MEP/DOXP, JA biosynthesis	Shi et al. (2015)
Adventitious root formation	Plant hormonal signal transduction, secondary metabolism, cell wall organization, glutathione metabolism	Wei et al. (2019)
Cold stress	Carbohydrate metabolism and calcium signaling pathway	Wang, Zhao, and Ma (2013)

AMT, ammonium transporter gene; NRT, nitrate transporter gene; AQP, aquaporin protein gene; GOGAT, glutamine (Gln) synthetase gene; MEP/DOXP, non-mevalonate (2C-methyl-D-erythritol-4-phosphate) pathway.

Similarly, seasonal variation also plays vital role in the tea quality and the overall yield. Kumar et al. (2016) studied the molecular basis of seasonal variation and identified the role of catechins and caffeine pathways in tea plants. Phytohormone metabolism, transcriptional regulation, and epigenetic control have been critical regulators of development and seasonal dormancy in tea plants. ABA biosynthesis-related genes are upregulated in phytohormone metabolism, and ABA catalyzes are downregulated during the dormant period (Kumar et al., 2016). There is a significant decrease of GA expression in tea during dormancy. Both GA20_{ox} and GA2_{ox} play a significant role in modulating the GA level in tea plants. The DELLA protein usually expresses highly during the dormancy, a potential indicator for quality leaf plucking intervals. During dormancy the responses of ABA and GA metabolism are known to be opposite. Pathways and genes involved in different stress management of tea plants are given in Table 22.3.

Alternative splicing (AS) is the posttranscriptional regulatory phenomenon that plays a significant role in generating multiple isoforms of the pre-mRNA transcripts and creates a diversity of the transcripts and proteomes (Mahadani & Hazra, 2021). An increasing number of studies have reported the role of alternative splicing events under different stresses and development stages in tea. AS events are mainly categorized into four common types, intron retention, exon skipping (ES), alternative 3' splice site (A3SS), and alternative 5' splice site (A5SS). Intron retention is the most common form of alternative splicing, followed by A3SS, A5SS, and ES. In tea an alternative splicing event is not only tissue-specific but also influences the flavonoid pathways. Major steps in the identification of alternative splicing events are mentioned in Table 22.4.

22.5 Discovery of single-nucleotide polymorphism

Single-nucleotide polymorphisms (SNPs) are widely used as an important molecular marker in plant genetic research and breeding. Due to advancements in sequencing technology, large numbers of genome-wide SNPs have been discovered by whole-genome sequencing or resequencing in nonmodel crops for linkage mapping, population structure and association studies, marker-assisted plant breeding, and functional genomics. SNP discovery in tea plants is challenging due to the complexity of the genome and lower levels of heterozygosity. However, the advancement of NGS-based software, pipeline, and the availability of standard reference genome, more SNPs discovery studies are reported from tea in recent years. The SNPs and indels of variant sites are identified by aligning the sequenced fragments with the latest reference genome. Large-scale SNPs and indels discoveries in tea plant are given in Table 22.5. These variant sites, especially tightly linked to the phenotypic expression or trait, are essential for the functional research and genomic-assisted breeding of tea trees. Validation of large numbers of SNPs is a major challenge for their successful implementation in tea breeding. Genome-wide indels are considered the third-generation molecular markers in plants due to their high polymorphism and reproducibility. It is expected that indels would emerge as a potential molecular marker in tea plants in the recent future.

TABLE 22.4 List of major studies of alternative splicing event identification from tea.

Bioprojects	Study area	Effect of alternative splicing event in tea	References	Platform
SRA: PRJNA524419	Anthocyanin biosynthesis pathway	<ul style="list-style-type: none"> 98 key genes undergone AS in anthocyanin biosynthesis pathways. <i>PAL2</i>, <i>C4H1</i>, <i>FLS1</i>, <i>CCR2</i>, <i>UDP75L122</i> and <i>MYB113-1</i> are major AS transcripts for regulating anthocyanin biosynthesis. 	Chen, L. et al. (2020)	PacBio RSII
SRA: PRJNA545401	Drought and heat stress	<ul style="list-style-type: none"> AS extensively triggered during drought and heat stress. ~48% of the genes in tea genome were differential spliced. 	Ding et al. (2020)	HiSeq 2500
SRA: PRJNA387105	Cold acclimation	<ul style="list-style-type: none"> AS event increase rapidly during cold and significantly decrease after deacclimation. AS genes mainly relate to the oxidoreductase activity and sugar metabolism pathways during cold acclimation. 	Li, Y. et al. (2020)	HiSeq X Ten
SRA: PRJNA274203	Different tissues	<ul style="list-style-type: none"> Intron retentions >Alternative 5' splice site >Exon skipping > Alternative 3' splice site. Regulate flavonoid pathways. 	Zhu et al. (2018)	HiSeq 2000 and PacBio RSII

TABLE 22.5 List of large-scale single-nucleotide polymorphism discovery studies in tea plants.

Area of research	No. of single-nucleotide polymorphism (SNP)	No. of indels	Strategies of SNP discovery	Reference
Varietal identification of tea (<i>Camellia sinensis</i>)	60	–	Mining from EST	Fang et al. (2014)
Large-scale SNP discovery and genotyping for construction a high-density genetic map of tea	6042	–	SLAF-Seq	Ma et al. (2015)
Genetic divergence between <i>C. sinensis</i> and its wild relatives	15444	–	RAD-Seq	Yang et al. (2016)
Genome-wide SNP detection in Darjeeling tea	54206	–	ddRAD-Seq	Hazra et al. (2020)
Analyses of SNPs identified by ddRAD-Seq reveal genetic structure	1269648	–	ddRAD-Seq	Yamashita et al. (2019)
Genetic diversity, linkage disequilibrium, and population structure analysis of the tea plant (<i>C. sinensis</i>)	79016	–	GBS	Niu et al. (2019)
Characterization of genome-wide genetic variations between two varieties of tea plant	7511731	255218	Genome-wide comparison	Liu et al. (2019)
Development of core collections for Guizhou tea genetic resources and GWAS of life size	30282	–	GBS	Niu et al. (2020)
Genetic diversity and adaptive evolution between two varieties of <i>C. sinensis</i>	18,903,625	7,314,133	Whole-genome resequencing	An et al. (2020)

GWAS, genome-wide association study; EST, expressed sequence tag; GBS, Genotyping by sequencing.

22.6 Conclusion

In the last decade the availability of tea genomic resources has increased considerably and opened many opportunities to decode this genomic information. This has resulted in the generation of big genomic data, which needs to be stored appropriately and categorized for further use. Few specialized databases on tea plants have been developed to help the tea research community. To date, more than eight whole-genome sequences of tea plants from different cultivars are publicly available. The selection of reference genome is vital for conducting any fruitful experiment in tea plants; although all the published genome has its own limitations and advantages. Huge SNP information has also been generated from tea plants, but reports of genomic-assisted breeding programs are very scarce compared to other vital crops. We hope that there will be more emphasis on genomics-assisted breeding of tea plants in the near future.

References

- Ahmad, M. Z., Li, P., She, G., Xia, E., Benedito, V. A., Wan, X. C., et al. (2020). Genome-wide analysis of serine carboxypeptidase-like acyltransferase gene family for evolution and characterization of enzymes involved in the biosynthesis of galloylated catechins in the tea plant (*Camellia sinensis*). *Frontiers in Plant Science*, *11*, 848.
- An, Y., Mi, X., Zhao, S., Guo, R., Xia, X., Liu, S., et al. (2020). Revealing distinctions in genetic diversity and adaptive evolution between two varieties of *Camellia sinensis* by whole-genome resequencing. *Frontiers in Plant Science*, *11*(1861).
- Board, T., (2019). *Tea Board of India 65th Annual Report 2018-19*.
- Cao, D., Liu, Y., Ma, L., Jin, X., Guo, G., Tan, R., et al. (2018). Transcriptome analysis of differentially expressed genes involved in selenium accumulation in tea plant (*Camellia sinensis*). *PLoS One*, *13*(6), e0197506.
- Chacko, S. M., Thambi, P. T., Kuttan, R., & Nishigaki, I. (2010). Beneficial effects of green tea: A literature review. *Chinese Medicine*, *5*(1), 1–9.
- Chatterjee, A., Paul, A., Unnati, G. M., Rajput, R., Biswas, T., Kar, T., et al. (2020). MAPK cascade gene family in *Camellia sinensis*: In-silico identification, expression profiles and regulatory network analysis. *BMC Genomics*, *21*(1), 613.

- Chen, J. D., Zheng, C., Ma, J. Q., Jiang, C. K., Ercisli, S., Yao, M. Z., et al. (2020). The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Horticulture Research*, 7, 63.
- Chen, L., Shi, X., Nian, B., Duan, S., Jiang, B., Wang, X., et al. (2020). Alternative splicing regulation of anthocyanin biosynthesis in *Camellia sinensis* var. *assamica* unveiled by PacBio Iso-Seq. *G3 (Bethesda)*, 10(8), 2713–2723.
- Ding, Y., Wang, Y., Qiu, C., Qian, W., Xie, H., & Ding, Z. (2020). Alternative splicing in tea plants was extensively triggered by drought, heat and their combined stresses. *PeerJ*, 8, e8258.
- Duan, Y., Zhu, X., Shen, J., Xing, H., Zou, Z., Ma, Y., et al. (2020). Genome-wide identification, characterization and expression analysis of the amino acid permease gene family in tea plants (*Camellia sinensis*). *Genomics*, 112(4), 2866–2874.
- Dubey, H., Rawal, H. C., Rohilla, M., Lama, U., Kumar, P. M., Bandyopadhyay, T., et al. (2020). TeaMiD: A comprehensive database of simple sequence repeat markers of tea. *Database (Oxford)*, 2020.
- Fang, W. P., Meinhardt, L. W., Tan, H. W., Zhou, L., Mischke, S., & Zhang, D. (2014). Varietal identification of tea (*Camellia sinensis*) using nano-fluidic array of single nucleotide polymorphism (SNP) markers. *Horticulture Research*, 1, 14035.
- Garg, R., & Jain, M. (2013). RNA-Seq for transcriptome analysis in non-model plants. *Methods in Molecular Biology*, 1069, 43–58.
- Guo, F., Guo, Y., Wang, P., Wang, Y., & Ni, D. (2017). Transcriptional profiling of catechins biosynthesis genes during tea plant leaf development. *Planta*, 246(6), 1139–1152.
- Guo, Y., Qiao, D., Yang, C., Chen, J., Li, Y., Liang, S., et al. (2020). Genome-wide identification and expression analysis of SABATH methyltransferases in tea plant (*Camellia sinensis*): Insights into their roles in plant defense responses. *Plant Signaling & Behavior*, 15(10), 1804684.
- Hao, X., Yang, Y., Yue, C., Wang, L., Horvath, D. P., & Wang, X. (2017). Comprehensive transcriptome analyses reveal differential gene expression profiles of *Camellia sinensis* axillary buds at para-, endo-, ecodormancy, and bud flush stages. *Frontiers in Plant Science*, 8, 553.
- Hazra, A., Kumar, R., Sengupta, C., & Das, S. (2020). Genome-wide SNP discovery from Darjeeling tea cultivars-their functional impacts and application toward population structure and trait associations. *Genomics*, 113(1), 66–78.
- Higdon, J. V., & Frei, B. (2003). Tea catechins and polyphenols: Health effects, metabolism, and antioxidant functions. *Critical Reviews in Food Science and Nutrition*, 43(1), 89–143.
- Hu, Z., Ban, Q., Hao, J., Zhu, X., Cheng, Y., Mao, J., et al. (2020). Genome-wide characterization of the C-repeat binding factor (CBF) gene family involved in the response to abiotic stresses in tea plant (*Camellia sinensis*). *Frontiers in Plant Science*, 11, 921.
- Jayaswall, K., Mahajan, P., Singh, G., Parmar, R., Seth, R., Raina, A., et al. (2016). Transcriptome analysis reveals candidate genes involved in blister blight defense in tea (*Camellia sinensis* (L) Kuntze). *Scientific Reports*, 6, 30412.
- Kumar, A., Chawla, V., Sharma, E., Mahajan, P., Shankar, R., & Yadav, S. K. (2016). Comparative transcriptome analysis of Chinary, Assamica and Cambod tea (*Camellia sinensis*) types during development and seasonal variation using RNA-seq technology. *Scientific Reports*, 6, 37244.
- Li, C. F., Zhu, Y., Yu, Y., Zhao, Q. Y., Wang, S. J., Wang, X. C., et al. (2015). Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (*Camellia sinensis*). *BMC Genomics*, 16, 560. Available from <https://doi.org/10.1186/s12864-015-1773-0>.
- Li, M., Lu, J., Tao, M., Li, M., Yang, H., Xia, E. H., et al. (2020). Genome-wide identification of seven polyamine oxidase genes in *Camellia sinensis* (L.) and their expression patterns under various abiotic stresses. *Frontiers in Plant Science*, 11, 544933.
- Li, W., Xiang, F., Zhong, M., Zhou, L., Liu, H., Li, S., et al. (2017). Transcriptome and metabolite analysis identifies nitrogen utilization genes in tea plant (*Camellia sinensis*). *Scientific Reports*, 7(1), 1693.
- Li, Y., Huang, J., Song, X., Zhang, Z., Jiang, Y., Zhu, Y., et al. (2017). An RNA-Seq transcriptome analysis revealing novel insights into aluminum tolerance and accumulation in tea plant. *Planta*, 246(1), 91–103.
- Li, Y., Wang, X., Ban, Q., Zhu, X., Jiang, C., Wei, C., et al. (2019). Comparative transcriptomic analysis reveals gene expression associated with cold adaptation in the tea plant *Camellia sinensis*. *BMC Genomics*, 20(1), 624.
- Li, Y., Mi, X., Zhao, S., Zhu, J., Guo, R., Xia, X., et al. (2020). Comprehensive profiling of alternative splicing landscape during cold acclimation in tea plant. *BMC Genomics*, 21(1), 65.
- Liu, S., An, Y., Tong, W., Qin, X., Samarina, L., Guo, R., et al. (2019). Characterization of genome-wide genetic variations between two varieties of tea plant (*Camellia sinensis*) and development of InDel markers for genetic research. *BMC Genomics*, 20(1), 935.
- Ma, J. Q., Huang, L., Ma, C. L., Jin, J. Q., Li, C. F., Wang, R. K., et al. (2015). Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). *PLoS One*, 10(6), e0128798.
- Mahadani, P., & Hazra, A. (2021). Expression and splicing dynamics of WRKY family genes along physiological exigencies of tea plant (*Camellia sinensis*). *Biologia*, 76, 2491–2499. Available from <https://doi.org/10.1007/s11756-021-00784-z>.
- Mondal, T. K., Rawal, H. C., Bera, B., Kumar, P. M., Choubey, M., Saha G., et al., (2019). Draft genome sequence of a popular Indian tea genotype TV-1 [*Camellia assamica* L. (O). Kunze]. *bioRxiv*. 762161.
- Mukhopadhyay, M., Mondal, T. K., & Chand, P. K. (2016). Biotechnological advances in tea (*Camellia sinensis* [L.] O. Kuntze): a review. *Plant Cell Reports*, 35(2), 255–287.
- Niu, S., Song, Q., Koiwa, H., Qiao, D., Zhao, D., Chen, Z., et al. (2019). Genetic diversity, linkage disequilibrium, and population structure analysis of the tea plant (*Camellia sinensis*) from an origin center, Guizhou plateau, using genome-wide SNPs developed by genotyping-by-sequencing. *BMC Plant Biology*, 19(1), 328.
- Niu, S., Koiwa, H., Song, Q., Qiao, D., Chen, J., Zhao, D., et al. (2020). Development of core-collections for Guizhou tea genetic resources and GWAS of leaf size using SNP developed by genotyping-by-sequencing. *PeerJ*, 8(e8572).
- Qiao, D., Yang, C., Chen, J., Guo, Y., Li, Y., Niu, S., et al. (2019). Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (*Camellia sinensis*). *Scientific Reports*, 9(1), 2709.

- Shen, J., Zou, Z., Xing, H., Duan, Y., Zhu, X., Ma, Y., et al. (2020). Genome-wide analysis reveals stress and hormone responsive patterns of JAZ family genes in *Camellia sinensis*. *International Journal of Molecular Sciences*, 21(7).
- Shi, C. Y., Yang, H., Wei, C. L., Yu, O., Zhang, Z. Z., Jiang, C. J., et al. (2011). Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics*, 12, 131.
- Shi, J., Ma, C., Qi, D., Lv, H., Yang, T., Peng, Q., et al. (2015). Transcriptional responses and flavor volatiles biosynthesis in methyl jasmonate-treated tea leaves. *BMC Plant Biology*, 15, 233.
- Wang, P., Yue, C., Chen, D., Zheng, Y., Zhang, Q., Yang, J., et al. (2019). Genome-wide identification of WRKY family genes and their response to abiotic stresses in tea plant (*Camellia sinensis*). *Genes Genomics*, 41(1), 17–33.
- Wang, X., Feng, H., Chang, Y., Ma, C., Wang, L., Hao, X., et al. (2020). Population sequencing enhances understanding of tea plant evolution. *Nature Communications*, 11(1), 4447.
- Wang, X. C., Zhao, Q. Y., Ma, C. L., Zhang, Z. H., Cao, H. L., Kong, Y. M., et al. (2013). Global transcriptome profiles of *Camellia sinensis* during cold acclimation. *BMC Genomics*, 14, 415.
- Wang, Y., Lu, Q., Xiong, F., Hao, X., Wang, L., Zheng, M., et al. (2020). Genome-wide identification, characterization, and expression analysis of nucleotide-binding leucine-rich repeats gene family under environmental stresses in tea (*Camellia sinensis*). *Genomics*, 112(2), 1351–1362.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1), 57–63.
- Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., et al. (2018). Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proceedings of the National Academy of Sciences of the United States of America*, 115(18), E4151–E4158.
- Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., et al. (2019). Auxin-induced adventitious root formation in nodal cuttings of *Camellia sinensis*. *International Journal of Molecular Sciences*, 20(19), 4817.
- Xia, E., Li, F., Tong, W., Yang, H., Wang, S., Zhao, J., et al. (2019). The tea plant reference genome and improved gene annotation using long-read and paired-end sequencing data. *Scientific Data*, 15, 122.
- Xia, E., Tong, W., Hou, Y., An, Y., Chen, L., Wu, Q., et al. (2020). The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Molecular Plant*, 13(7), 1013–1026.
- Xia, E. H., Zhang, H. B., Sheng, J., Li, K., Zhang, Q. J., Kim, C., et al. (2017). The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Molecular Plant*, 10(6), 866–877.
- Xia, E. H., Li, F. D., Tong, W., Li, P. H., Wu, Q., Zhao, H. J., et al. (2019). Tea plant information archive: A comprehensive genomics and bioinformatics platform for tea plant. *Plant Biotechnology Journal*, 17(10), 1938–1953.
- Xing, A., Ma, Y., Wu, Z., Nong, S., Zhu, J., Sun, H., et al. (2020). Genome-wide identification and expression analysis of the CLC superfamily genes in tea plants (*Camellia sinensis*). *Functional & Integrative Genomics*, 20(4), 497–508.
- Xu, P., Zhang, X., Su, H., Liu, X., Wang, Y., & Hong, G. (2020). Genome-wide analysis of PYL-PP2C-SnRK2s family in *Camellia sinensis*. *Bioengineered*, 11(1), 103–115.
- Yamashita, H., Katai, H., Kawaguchi, L., Nagano, A. J., Nakamura, Y., Morita, A., et al. (2019). Analyses of single nucleotide polymorphisms identified by ddRAD-seq reveal genetic structure of tea germplasm and Japanese landraces for tea breeding. *PLoS One*, 14(8), e0220981.
- Yang, H., Wei, C. L., Liu, H. W., Wu, J. L., Li, Z. G., Zhang, L., et al. (2016). Genetic divergence between *Camellia sinensis* and its wild relatives revealed via genome-wide SNPs from RAD sequencing. *PLoS One*, 11(3), e0151424.
- Yu, Q., Li, C., Zhang, J., Tian, Y., Wang, H., Zhang, Y., et al. (2020). Genome-wide identification and expression analysis of the Dof gene family under drought stress in tea (*Camellia sinensis*). *PeerJ*, 8, e9269.
- Zhang, C. C., Wang, L. Y., Wei, K., Wu, L. Y., Li, H. L., Zhang, F., et al. (2016). Transcriptome analysis reveals self-incompatibility in the tea plant (*Camellia sinensis*) might be under gametophytic control. *BMC Genomics*, 17(17), 359.
- Zhang, D., Han, Z., Li, J., Qin, H., Zhou, L., Wang, Y., et al. (2020). Genome-wide analysis of the SBP-box gene family transcription factors and their responses to abiotic stresses in tea (*Camellia sinensis*). *Genomics*, 112(3), 2194–2202.
- Zhang, Q., Cai, M., Yu, X., Wang, L., Guo, C., Ming, R., et al. (2017). Transcriptome dynamics of *Camellia sinensis* in response to continuous salinity and drought stress. *Tree Genetics & Genomes*, 13, 78.
- Zhang, Q. J., Li, W., Li, K., Nan, H., Shi, C., Zhang, Y., et al. (2020). The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. *Molecular Plant*, 13(7), 935–938.
- Zhang, R., Ma, Y., Hu, X., Chen, Y., He, X., Wang, P., et al. (2020). TeaCoN: A database of gene co-expression network for tea plant (*Camellia sinensis*). *BMC Genomics*, 21(1), 461.
- Zhang, W., Zhang, Y., Qiu, H., Guo, Y., Wan, H., Zhang, X., et al. (2020). Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nature Communications*, 11. Available from <https://doi.org/10.1038/s41467-020-17498-6>.
- Zhang, X., Xu, W., Ni, D., Wang, M., & Guo, G. (2020a). Genome-wide characterization of tea plant (*Camellia sinensis*) Hsf transcription factor family and role of CsHsfA2 in heat tolerance. *BMC Plant Biology*, 20(1), 244.
- Zhang, X., Li, Q., Xu, W., Zhao, H., Guo, F., Wang, P., et al. (2020b). Identification of MTP gene family in tea plant (*Camellia sinensis* L.) and characterization of CsMTP8.2 in manganese toxicity. *Ecotoxicology and Environmental Safety*, 202, 110904.
- Zhou, Q., Zhao, S., Zhu, J., Li, F., Tong, W., Liu, S., et al. (2020). Transcriptomic analyses reveal a systemic defense role of the uninfested adjacent leaf in tea plant (*Camellia sinensis*) attacked by tea geometrids (*Ectropis obliqua*). *Genomics*, 112(5), 3658–3667.
- Zhu, J., Wang, X., Xu, Q., Zhao, S., Tai, Y., & Wei, C. (2018). Global dissection of alternative splicing uncovers transcriptional diversity in tissues and associates with the flavonoid pathway in tea plant (*Camellia sinensis*). *BMC Plant Biology*, 18(1), 266.

This page intentionally left blank

Next-generation sequencing and viroid research

Sunny Dhir, Asha Rani and Narayan Rishi

Amity Institute of Virology & Immunology, Amity University, Noida, Uttar Pradesh, India

23.1 Introduction

Viroids are single-stranded circular RNAs, highly structured and compact due to presence of self-complementarity among their nucleotides. They have small-size RNA in the range of 246–401 nucleotides and have ability to cause disease that can have mild-to-severe symptoms on respective host plants. They are around 10 times smaller than smallest known RNA virus. Viroids do not have any protein coding capacity and are entirely dependent upon host factors for their infectivity and life cycle (Flores, Hernández, Alba, Daròs, & Serio, 2005). It is fascinating that in viroid biology the RNA sequence that is not translated can cause disease. Viroids were first discovered in spindle tuber disease of potato (Allison, Simon, & Maliga, 1996) and since then more than 32 species of viroids are known. This could be possible due to advancements in gene sequencing technologies such as next-generation sequencing (NGS) (Adkar-Purushothama & Perreault, 2020). Viroids infect higher plants and cause diseases that result in huge economic losses every year. Yield losses up to 100% have been reported due to viroid infection (Jones, Baizan-Edge, MacFarlane, & Torrance, 2017). They affect wide variety of crops such as potato, cucumber, hop, coconut, tomato and grapevine, subtropical and temperate fruit trees such as avocado, apple, peach, pear, citrus, and plum. Ornamental plants such as coleus and chrysanthemum are known to be infected with viroids. The mode of transmission can be mechanical, through seeds or pollens. Transmission through aphids has also been reported in case of *Tomato planta macho viroid* (TPMVd) but within specific ecological conditions (Matsuura, Matsushita, Kozuka, Shimizu, & Tsuda, 2010). Active transmission of *Apple scar skin viroid* (ASSVd) by the whitefly *Trialeurodes vaporariorum* is also reported. Electrophoretic mobility shift assay and northwestern hybridization assays were used for the determination of phloem proteins with ASSVd for its efficient transmission (Walia, Dhir, Zaidi, & Hallan, 2015). The most efficient mode for transmission is through vegetative propagation using infected material. This might be the reason for the presence of mixture of viroids in grapevine and citrus plants which are propagated in the same way, that is, using vegetative parts of infected plants.

Viroids are divided into two families based upon the structure of the RNA and region of their replication, the *Avsunviroidae* and the *Pospiviroidae* (Wang, 2021). The latter have rod-shaped secondary RNA structure and replicate in the nucleus, while the former has hammerhead-like structure and replicate in chloroplasts (Allison et al., 1996). Members of the family *Pospiviroidae* replicate via asymmetric rolling-circle amplification, while those of *Avsunviroidae* by symmetric rolling-circle amplification. Viroid families, genus, along with species are described in Table 23.1.

Viroid replication biology depicts the existence of (–) polarity RNA sequences along with that of (+) as intermediates in the replication process. The preferential accumulation of these strands was observed in chloroplast, apart from nucleus, which suggests the role of chloroplast in replication (Moreno et al., 2019). Viroids are known to have several sequence variants and may lead to attenuation of disease symptoms. For instance, *Apple scar skin viroid* causes fruit scarring and its variant, dapple apple, causes dappling of apple fruits (Allison et al., 1996). There are only few reports on the biology of the characterized viroids and their variants (Walia, Dhir, Bhadoria, Hallan, & Zaidi, 2012). Viroids do not encode for any proteins, still they are able to evade host's defense mechanism, replicate, and propagate in the plant. Due to their highly base paired structures and RNA–RNA mode of replication, viroids are inducers as well as

TABLE 23.1 Viroid genera, families, and species.

Family	Genus	Viroids
<i>Avsunviroidae</i>	1. <i>Avsunviroid</i> 2. <i>Pelamoviroid</i> 3. <i>Elaviroid</i>	1. <i>Avocado sunblotch viroid</i> 2. <i>Chrysanthemum chlorotic mottle viroid</i> , <i>Peach latent mosaic viroid</i> 3. <i>Eggplant latent viroid</i> , <i>Grapevine hammerhead viroid-like RNA</i> , <i>Apple hammerhead viroid-like RNA</i>
<i>Pospiviroidae</i>	1. <i>Pospiviroid</i> 2. <i>Hostuviroid</i> 3. <i>Cocadviroid</i> 4. <i>Apscaviroid</i> 5. <i>Coleviroid</i>	1. <i>Potato spindle tuber viroid</i> , <i>Tomato apical stunt viroid</i> , <i>Tomato chlorotic dwarf viroid</i> , <i>Tomato planta macho viroid</i> , <i>Columnnea latent viroid</i> , <i>Citrus exocortis viroid</i> , <i>Chrysanthemum stunt viroid</i> , <i>Pepper chat fruit viroid</i> , <i>Iresine viroid I</i> , <i>Portulaca latent viroid</i> 2. <i>Hop stunt viroid</i> , <i>Dahlia latent viroid</i> 3. <i>Coconut cadang-cadang viroid</i> , <i>Coconut tinangaja viroid</i> , <i>Citrus bark cracking viroid</i> , <i>Hop latent viroid</i> 4. <i>Apple scar skin viroid</i> , <i>Apple dimple fruit viroid</i> , <i>Pear blister canker viroid</i> , <i>Citrus bent leaf viroid</i> , <i>Citrus dwarfing viroid</i> , <i>Citrus viroid V</i> , <i>Citrus viroid VI</i> , <i>Citrus viroid OS</i> , <i>Australian grapevine viroid</i> , <i>Grapevine yellow speckle viroid 1</i> , <i>Grapevine yellow speckle viroid 2</i> , <i>Apple fruit crinkle viroid</i> , <i>Grapevine yellow speckle viroid 3</i> , <i>Grapevine latent viroid</i> , <i>Persimmon latent viroid</i> , <i>Persimmon viroid 2</i> 5. <i>Coleus blumei viroid 1</i> , <i>Coleus blumei viroid 2</i> , <i>Coleus blumei viroid 3</i> , <i>Coleus blumei viroid 4</i> , <i>Coleus blumei viroid 5</i> , <i>Coleus blumei viroid 6</i>

targets of RNA silencing, a defense mechanism of the host. RNA silencing leads to the generation of small RNAs that are taken up by Argonaut proteins to inactivate messenger RNAs, leading to disease. However, the mechanism by which members of the *Pospiviroidae*-derived small RNAs causing disease is still elusive (Flores, Navarro, Delgado, Serra, & Di Serio, 2020).

Viroids cause plethora of symptoms on infected plants. Some viroids destroy whole cultivar, while some show chlorosis, chlorotic spots sometime covering the whole blade, epinasty, rugosity, pitting, internode shortening and stem dwarfing, scaling, cracking, canker on bark, stunting, broken lines on petals (flowers) discolorations and skin deformations, suture cracking on fruits, enlarged stones (seeds), delays in foliation, flowering, ripening, and growing pattern of mature trees (Flores et al., 2005). The symptoms can be organ specific or can be present all over. Few viroids may show mild-to-no symptoms as in case of infection in wild plants. High light intensity and high temperature in contrast to that of viruses helps in the expression of symptoms. Thus thermotherapy is not successful in eradication of viroids. However, cross protection is observed in viroid infection in which the plant infected with a mild strain of a viroid and protects the host against infection with the severe strain of the same viroid. The characteristic symptoms are suppressed for some time (Flores et al., 2005).

With increased viroid disease incidence and heavy crop losses, viroids have become important pathogens. Identification of the viroid associated with a disease was like finding a needle in a haystack because of the small-sized genomic RNA. With recent developments in sequencing technologies, viroid research areas have exponentially progressed. NGS provides highly efficient, robust amplification, and sequencing platform corroborated with bioinformatics with which viroid discovery and its association with disease has become easier. For more than a decade, the utilization of NGS has led to increased discovery of viroids as well as proven potential in deciphering other mechanisms in viroid research. Here, in this book chapter, we have tried to put together the role of NGS in viroid discovery as well as its role in understanding the viroid RNA biology, including mutation analyses and pathogenicity. The use of various bioinformatic tools in NGS is also discussed.

23.2 Next-generation sequencing technology

NGS has been used since decades for various biotechnological applications related to genomics, transcriptomics, and proteomics. Its use in diagnostics was reported earlier in year 2007 in clinical virology and since then the number of studies appeared as viral diagnostics as well as in understanding the infection processes (Grada & Weinbrecht, 2013). With the development in sequencing techniques and NGS, there has been increase in discovery of new viroids as well as new hosts. Without any prior information regarding any target, NGS can produce results specific to the target strain (Adams & Fox, 2016). “Next generation” is the term given to the development in sequencing technology to the next

level (Slatko, Gardner, & Ausubel, 2018). The development in sequencing technology differs in the method of sequencing as here glass slides are used on which millions of template DNA strands are bind at a discrete position. There becomes single modified base that extends the template. These modified bases are labeled with fluorescent dye and microscope captures the image reflecting both the position as well as the intensity of the fluorescent color. The unique step involves the conversion of modified bases to regular one and the imaging continues with extension of each nucleotide base on template strand. After several cycles, the colored map is obtained in the form of bases A, T, G, or C. The single template tells about the sequence of a particular length that is known as “read.” However, the initial steps are similar as used in initial sequencing methods such as Sanger’s sequencing method. But the restoration step makes a difference along with its high speed (Muzzey, Evans, & Lieber, 2015).

23.3 Impact of next-generation sequencing on viroid discovery

The first DNA sequencing approaches involved chemical methods, including 2D chromatography, Maxam–Gilbert, and Sanger sequencing. Further with the development of polymerase chain reaction (PCR), using good-quality enzymes and fluorescent automated DNA sequencing techniques provided more facts regarding viroids. Later in 2006 high-throughput sequencing methods led to the study of billions of DNA and RNA sequences and since then, NGS continuously has helped the researchers to explore various principles of viroid biology. Advancements made overtime in understanding viroid RNA biology along with sequencing techniques used are shown in Table 23.2 (Adkar-Purushothama & Perreault, 2020).

NGS can detect multiple (RNA and DNA) viruses infecting a single plant at a given time that can also help in unraveling disease antagonism or synergism mechanisms through transcriptomic approach that was earlier not possible. Being a sequence-independent technique, NGS can detect viruses and viroids which earlier remained undetectable using primitive molecular and serological methods (Jones et al., 2017). For instance, transcriptome data obtained using NGS of an infected grapevine cultivar from different tissues, namely, grain, skin, and seeds was compared against reference sequences of virus genomes. Sequence analyses yielded multiple infection with viroids and virus, namely, *Grapevine yellow speckle viroid 1*, *Grapevine pinot gris virus*, *Hop stunt viroid*, and *Grapevine leafroll-associated virus 2* as most prevalent. The outline for the complete process followed for the identification of mixed infection is shown in Fig. 23.1

TABLE 23.2 Various sequencing platforms used in deciphering viroid RNA biology (Kulski, 2016).

Sequencing platforms	Techniques	Viroid biology
First generation	<ul style="list-style-type: none"> • Two-dimensional chromatography and spectrophotometric procedures • Maxam and Gilbert • Sanger technique • Automated DNA sequencing (PCR technology, fluorescent dye) 	Biological tests and infection assays for viroid identification 2D fractionation technique for first viroid sequencing [PSTVd (1978) and CEVd] Broad classification of viroids into <i>Pospiviroidae</i> and <i>Avsunviroidae</i>
Second generation	<ul style="list-style-type: none"> • Shotgun sequencing (linkers/adapters) • Pyrosequencing (Roche 454 pyrosequencing by synthesis) • Illumina HiSeq and MiSeq sequencing (fluorescently labeled nucleotides) • Sequencing by Oligonucleotide Ligation and Detection, that is, SOLiD (annealing of probes to template and ligation) • DNA nanoball sequencing by BGI Retrovolocity (Nanoballs with DNA amplified on it are attached to an arrayed flow cell) • Ion Torrent (with microchips and sensors, nucleotides are incorporated as electronic signal) 	Detection of multiple viroids in a single assay Small-RNA sequencing Transcriptome analysis Whole-genome sequencing Genome-wide mutational analysis Characterization of genetic variations and impact of viroid infection on host cells and functions
Third generation	<ul style="list-style-type: none"> • Single Molecule Real Time, that is, SMRT sequencing method (template and DNA polymerase coupled on ultrawells and detection of nucleotide after each incorporation) • Helicos sequencing system (addition of polyA-tailed nucleotides to Oligo-DT) • Nanopore sequencing (conductivity changes as a nucleotide passes through nanopore) • Electron microscopy 	

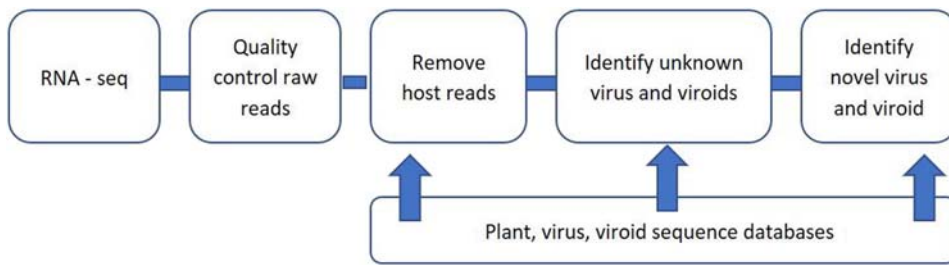


FIGURE 23.1 Outline of next-generation sequencing used in viroid discovery. Viroid-derived small RNAs are isolated (library preparation) and subjected to amplification and sequencing.

(Walia et al., 2015). The data obtained was corroborated by RT-PCR for the presence of viroids and viruses in the tested samples. Detection and identification of viroids using NGS in various host plants is listed in Table 23.3 (Barba & Hadidi, 2017; Gucek et al., 2017; Hadidi, 2019; Jakše, Radišek, Pokorn, Matoušek, & Javornik, 2015) (Fig. 23.2).

23.4 Role of next-generation sequencing in unraveling viroid RNA biology

23.4.1 Characterization of viroid sequence variants

For the detection of *Hop stunt viroid* (HSVd), 2D-PAGE and sequential PAGE were used, but sequencing was still in need for complete detection. The problem was solved with the development of hybridization-based methods. But with the limitation of sensitivity, RT-PCR became one of the most sensitive techniques. However, high-throughput sequencing allowed detection of multiple viroids/variants using small-RNA deep sequencing.

Viroids upon infection leads to the generation of viroid variants, such as viruses, that can induce an array of symptoms on host plants. They use host RNA polymerase that lacks proofreading activity and hence leads to generation of different variants also known as “quasispecies.” Sequence variants were reported first in case of CEVd followed by PSTVd that has mild, moderate, and severe sequence variants responsible for inducing similar effect on tomato plants (Adkar-Purushothama, Sano, & Perreault, 2018). Four different sequence variants were also reported in ASSVd infection in cucumber and apple using Single Strand Conformation Polymorphism corroborated by sequencing (Walia, Dhir, Ram, Zaidi, & Hallan, 2014). High-fidelity ultradeep sequencing revealed high mutation frequencies in ELVd and PSTVd infecting eggplant (López-Carrasco et al., 2017). Quasispecies generated during PSTVd infection was also studied using deep sequencing of viroid small RNAs that also led to identification of strand-specific mutations and revealed hotspots for mutations (Brass, Owens, Matoušek, & Steger, 2017). A similar study identified the regions on viroid genome that were favored for mutations and their effect on viroid secondary structure and small-RNA generation was revealed using NGS (Adkar-Purushothama, Bolduc, Bru, & Perreault, 2020). The development of NGS platform led to characterization of multiple viroid/variants and contributed to disease outcome due to presence of different sequence variants. In one such study, genetic diversity of PLMVd and PSTVd infected with single-sequence variant in both cases was evaluated. PLMVd, member of the family Avsunviroidae, revealed variant sequences with mutations at 50% of the viroid genome while mutations were lower in sequence variants of PSTVd (Glouzon, Bolduc, Wang, Najmanovich, & Perreault, 2014).

23.4.2 Viroid pathogenesis

Pathogenesis in viroid infection has been associated with generation of viroid-derived small RNAs that lead to symptom expression. The small RNAs bind to complementary endogenous RNAs and inhibit their expression. Such transcriptional changes play a major role in viroid pathogenesis. NGS helped in deciphering the processes involved in viroid pathogenesis. Different pathways associated with different cell organelles were studied to analyze the viroid infection process. In a study on two cultivars of tomato having infection of mild and severe strains of PSTVd, the genes related to chloroplast were downregulated and other genes related to nucleus, cell wall, ribosome, etc. were upregulated in one of the cultivars (Visvader & Symons, 1983). In another study, the genes related to *brassinosteroids* synthesis were detected as when the sources are applied, the genes were upregulated (Owens, Tech, Shao, Sano, & Baker, 2012). Northern blot hybridization and sequencing revealed 21–24 nucleotide viroid-derived small RNAs and later the analysis of RNA guiding, the step, which was sequence specific, revealed that the silencing machinery is being operated by viroids (St-Pierre, Hassen, Thompson, & Perreault, 2009). This has also been evidenced in a study that involved generation of transgenics expressing viroid small RNAs verified the role of RNAi-based inhibition against PSTVd infection (Adkar-Purushothama et al., 2015).

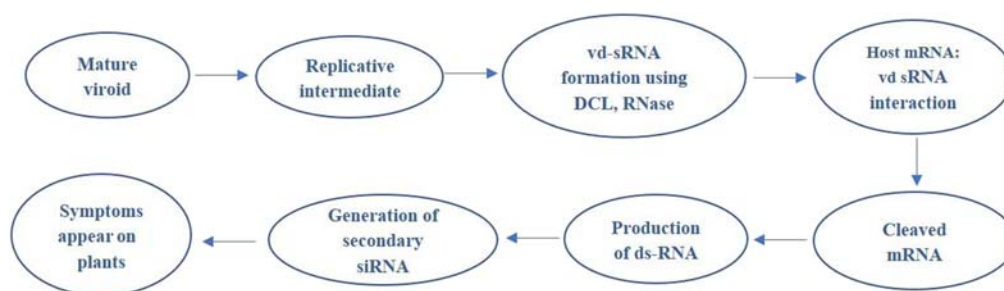
TABLE 23.3 List of viroid species identified in different hosts using next-generation sequencing (NGS).

Viroid	Host	NGS	Technique used for detection
<i>Apple dimple fruit viroid</i>	Apple, <i>Prunus</i> , pear, fig	Illumina Hi Scan SQ	SDS-PAGE, northern blot hybridization, RT-PCR (Shamloul, Faggioli, Keith, & Hadidi, 2002)
<i>Apple fruit crinkle viroid</i>	Apple, hop	Illumina HiSeq	2D-PAGE, SDS-PAGE, dot blot hybridization, multiplex RT-PCR (Koganezawa, 1985)
<i>Apple scar skin viroid</i>	Apple, pear, <i>Prunus</i> , sweet cherry, apricot, quince	Illumina HiSeq 2500	Dot blot hybridization, tissue blot hybridization, multiplex RT-PCR, RT-PCR-ELISA (Hadidi & Yang, 1990)
<i>Hop stunt viroid</i>	Cucumber, grapevine, orange, plum, peach, pear, apple, almond, apricot, hop	Illumina Genome Analyzer IIx	NGS, multiplex RT-qPCR, northern blot hybridization, SDS-PAGE (Sasaki & Shikata, 1977)
<i>Peach latent mosaic viroid</i>	<i>Prunus</i> , nectarine, peach, pear, apricot, quince	Illumina	Tissue dot blot hybridization, multiplex RT-qPCR, NGS, dot blot hybridization, RT-LAMP (Di Serio, Malfitano, Flores, & Randles, 1999)
<i>Tomato chlorotic dwarf viroid</i>	Tomato, <i>Petunia</i> spp.	Deep sequencing, Illumina	Southern blot hybridization, dot blot hybridization, multiplex RT-PCR, RT-qPCR (Matsushita, Kanda, Usugi, & Tsuda, 2008)
<i>Tomato apical stunt viroid</i>	Tomato, potato	Illumina	Northern blot hybridization, dot blot hybridization, microarray, multiplex RT-PCR (Antignus, Lachman, Pearlsman, Gofman, & Bar-Joseph, 2002)
<i>Tomato planta macho viroid</i>	Tomato	Illumina	RT-PCR, polyprobe dot blot hybridization, RT-qPCR (Verhoeven, Roenhorst, & Owens, 2011)
<i>Potato spindle tuber viroid</i>	Potato, tomato, <i>Dahlia</i> , <i>Petunia</i>	Illumina Genome Analyzer IIx	Dot blot hybridization, tissue blot hybridization, dot and print RT-PCR, RT-PCR-ELISA, RT-LAMP (Owens & Diener, 1981)
<i>Pepper chat fruit viroid</i>	Sweet pepper, tomato	NGS	R-PAGE, RT-PCR, RT-qPCR, polyprobe dot blot hybridization, multiplex RT-PCR (Botermans et al., 2020)
<i>Mexican papita viroid</i>	<i>Solanum cardiophyllum</i> (heartleaf nightshade), Tomato	Illumina	R-PAGE, RT-qPCR, polyprobe dot blot hybridization, RT-PCR, Multiplex RT-PCR (Martínez-Soriano et al., 1996)
<i>Columnnea latent viroid</i>	<i>Columnnea erythrophae</i> , tomato	NGS	PAGE, Polyprobe Dot Blot hybridization, RT-PCR, RT-qPCR (Hammond, Smith, & Diener, 1989)
<i>Chrysanthemum chlorotic mottle viroid</i>	<i>Chrysanthemum</i>	NGS	Micro tissue direct RT-PCR, ICAN, RT-PCR (Hosokawa, Matsushita, Uchida, & Yazawa, 2006)
<i>Avocado sunblotch viroid</i>	Avocado	NGS	RT-PCR (Schnell, Kuhn, Ronning, & Harkins, 1997)
<i>Citrus exocortis viroid</i>	Orange, grape, tomato	Illumina Genome Analyzer IIx	Dot blot hybridization, RT-PCR (de Noronha Fonseca, Marcellino, & Gander, 1996)
<i>Chrysanthemum stunt viroid</i>	<i>Petunia</i> , potato, tomato, <i>Chrysanthemum</i>	Roche 454 Y	RT-PCR, polyprobe dot blot hybridization, multiplex RT-LAMP (Mumford, Walsh, & Boonham, 2000)
<i>Coleus blumei viroids</i>	<i>Coleus</i> (<i>Plectranthus scutellarioides</i>)	NGS	PAGE, northern blot hybridization (Hou, Li, Wu, Jiang, & Sano, 2009)
<i>Hop latent viroid</i>	Hop (<i>Humulus lupulus</i>)	NGS	Dot blot hybridization, multiplex RT-PCR (Matoušek & Patzak, 2000)
<i>Citrus bark cracking viroid</i>	Grapefruit, hop	Illumina, transcriptome sequencing	NGS, RT-PCR (Owens, Sano, & Duran-Vila, 2012)

(Continued)

TABLE 23.3 (Continued)

Viroid	Host	NGS	Technique used for detection
<i>Coconut cadang-cadang viroid</i>	African oil palm, coconut palm, buri palm	NGS	2D-PAGE, RT-LAMP (Vadamalai, Hanold, Rezaian, & Randles, 2006)
<i>Pear blister canker viroid</i>	Pear, quince, apple	NGS	Multiplex RT-PCR-ELISA, multiplex RT-PCR (Hadidi & Yang, 1990)
<i>Grapevine yellow speckle viroid 1</i>	Grapevine	Illumina Genome Analyzer IIx	RT-PCR, dot blot hybridization (Teruo, Kobayashi, Ishiguro, & Motomura, 2000)
<i>Grapevine yellow speckle viroid</i>	Grapevine	Illumina	RT-PCR (Koltunow, Krake, Johnson, & Rezaian, 1989)
<i>Grapevine yellow speckle viroid 2</i>	Grapevine	Illumina	Multiplex RT-PCR, 2D-PAGE (Flores, Hernandez, Llacer, & Desvignes, 1991)
<i>Citrus viroid V</i>	Citrus spp.	NGS	RT-PCR, northern blot (Serra et al., 2008)
<i>Citrus dwarfing viroid</i>	Citrus, citron, orange, grapefruit	NGS	RT-PCR, northern blot hybridization (Malfitano, Barone, Duran-Vila, & Alioto, 2005)
<i>Citrus bent leaf viroid</i>	Citrus, Citron	NGS	Dot blot hybridization, RT-PCR (Zhang et al., 2014)
<i>Grapevine latent viroid</i>	Grapevine	Illumina HiSeq-2000	NGS (Fadda, Daròs, Fagoaga, Flores, & Durán-Vila, 2003)
<i>Eggplant latent viroid</i>	Eggplant	Illumina MiSeq machine sequencer	PAGE, northern blot hybridization (Fadda et al., 2003)

**FIGURE 23.2** A sequence of events occurs for the production of viroid derived small RNAs (vdsRNAs) during which the symptoms appear on plants and then the silencing mechanism operates due to the unusual double stranded RNA structure formation.

Viroids are inducer as well as targets of RNA silencing mechanism in host plants. Different groups have shown the role of viroid-derived small RNAs targeting host mRNAs and induce pathogenesis (Ramesh et al., 2020). Transgenic plants expressing a region from PSTVd implicated to target an endogenous mRNA induced abnormal phenotypes similar to PSTVd-infected plants (Adkar-Purushothama et al., 2015). High-throughput sequencing techniques for the analysis of viroid sRNAs and northern blot assay to detect mature forms of viroid were used for comparison between infected and healthy RNA samples from cucumber plant infected with HSVd (Martinez, Donaire, Llave, Pallas, & Gomez, 2010). The plus- and minus-sense RNAs were equally present. Similar findings were obtained in the case of *Grapevine yellow speckle viroid* but contrarily infection in case of CEVd and PSTVd revealed more amount of sense viroid RNA compared to antisense RNA. The complexity in generation of 21–24 nucleotide sRNAs (hallmark of RNA

silencing) species revealed highest complexity in case of 21 nucleotide long sRNAs rather than 22, 23, and 24-nt sized RNAs during HSVd infection (Martinez et al., 2010). Few other findings demonstrated the involvement of RNA silencing machinery like, when a PLMVd variant carrying a hair-pin sequence induced albinism in host plants which was due to targeting of endogenous mRNA encoding heat shock protein 90 (HSP90) and a PSTVd sRNA target the callose synthase gene in an artificial microRNA (miRNA) experiment (Zhang, Wu, Li, & Wu, 2015). Transcriptome data of infected plants when compared with normal plants revealed differences that might have been triggered by viroid infection. This study was carried out first by using microarray technologies and later RNA sequencing was used and showed changes in many processes such as photosynthesis and RNA regulation (Štajner et al., 2019). In case of HSVd infection, the transcriptomic study revealed differences in lipid metabolism, photosynthesis (depressed), expression of RNA-dependent RNA polymerase, etc. Besides, 2000 genes including protein metabolism, pigment metabolism, immune response in plants, and phytohormone signaling were also modulated (Mishra et al., 2018).

23.4.3 Mutational analyses of the viroids

Mutation-related studies have also been done using site-directed mutagenesis which somewhere has helped in identifying the regions of viroid related to degree of pathogenicity, its movement, and its replication. Some isolates of the same viroid showed different degree of severity in infection. Thus, on comparing the sequence in isolates and inducing site-directed mutations, the region related to pathogenicity was revealed. In one such studies, a change of single nucleotide in upper part of the central conserved region results in loss of infectivity in CEVd (Visvader, Forster, & Symons, 1985). An isolate of PSTVd from tomato inoculated in tobacco plant resulted in nucleotide substitution from “cytosine” to “uracil” at position 259 of the viroid for effective replication. Similarly, mutation from “uracil” to “adenine” at position 257 changed the PSTVd strain into a lethal strain when inoculated on tomato plant. Substitution at 257th position with “adenine” or “cytosine” has improved the viroid replication in tobacco plant (Qi & Ding, 2002; Qi & Ding, 2003). *Coleus blumei* viroid (CBVd) is seed transmissible and to identify nucleotide/nucleotides responsible, genome-wide mutants of CBVd were generated that revealed 25th nucleotide in loop five was responsible for viroid transmission through seeds (Tsushima & Sano, 2018).

PSTVd sequence variants were studied based on the mutation over a period of 1 week and 2 weeks, respectively. The variations were dominating after 1 week post inoculation and original sequence was found to be 25%. Two weeks post inoculation, the original sequence was 70% till the infection was there (Adkar-Purushothama et al., 2020). Genome-wide mutants were generated for PSTVd to generate a viroid genomic map with sequence characteristics involved in trafficking, replication, and pathogenicity (Zhong, Archual, Amin, & Ding, 2008). NGS has its own advantage in detection of sequence variants in purified isolates. The sequence analysis of full-length cDNA clones of CEVd and PSTVd isolates revealed the presence of sequence variants in the same host. The reason speculated was high copy error rate during replication or due to presence of multiple sequence variants during propagation (Visvader & Symons, 1985). It is possible that a particular naturally infected cultivar from viroids may contain a mixed infection of variants in the same host (Visvader & Symons, 1985). For instance, AFCVd transferred to tomato, cucumber, and hop, had incorporation of host-dependent sequence changes that appeared in naturally occurring other AFCVd isolates using small-RNA deep sequencing. It also indicated that the left-hand half of the viroid genome is critical for infection (Suzuki et al., 2017). High-throughput sequencing such as massively parallel sequencing helped the analysis of multiple sequences of viroid RNA molecules much faster than first-generation sequencing techniques.

23.5 Bioinformatic intervention in next-generation sequencing

Bioinformatics is the field of science that develops different methods and software tools to understand the complex biological data and address them from computational point of view. Various bioinformatic tools have been developed to interpret the data obtained through NGS. The main steps involved in sequence analysis were (1) accessing the data banks such as GenBank that has publicly available nucleotide sequences, (2) using appropriate tools to analyze the data such as FASTA, and (3) interpretation of the results in biological manner (Kamble & Khairkar, 2016). NGS has provided platforms for fast, efficient, and robust sequencing technology. But this advancement has been possible by corroboration and unprecedented updation in tools that process raw data. Bioinformatic tools allow the discovery of novel viruses and viroids using homology-dependent and homology-independent identification. After processing the raw data generated by NGS platform, there occurs sequence assembly of the preprocessed reads using various tools such as Velvet, Oases, and VCAKE (Mehmood, Sehar, & Ahmad, 2014). Similarly, the assembled reads are queried using the homology-dependent tools such as USEARCH, HHbits, and SearchSmallRNA that were developed to make assembly

of viral/viroid genomes in a much easier and reliable manner. Other than this, CLC Genomics Workbench, Geneious, Galaxy, etc. were developed to facilitate the use of NGS technologies as they provided user-friendly interfaces.

It becomes easy to identify a pathogen using sequence homology approach using database. But metagenomic approach makes it easier to identify a pathogen without any prior knowledge. Such programs included PFOR (Progressive Filtering of Overlapping Small RNAs) that includes filtering of overlapping sRNA sequences for the assembly of complete genome of viroids, as sRNA sequences are hallmark of viroid infection. Identification of *Grapevine latent viroid* (GLVd) was done by this homology-independent identification metagenomic approach. PFOR/PFOR2 is helpful in discovery of completely new Grapevine hammerhead viroid-like RNA and Apple hammerhead viroid-like RNA. The detection of viroid sequences is possible without ribosomal RNA depletion and the available bio-informatic tools do not provide a suitable platform for the identification of new viroid sequence. SLS developed as part of PFOR2 discover biologically active circular RNAs by deep sequencing of long RNAs (Wu, Ding, Zhang, & Zhu, 2015). It successfully assembled PSTVd genome from small RNAs sequenced from infected plant after rRNA depletion. Thus SLS-PFOR2 allowed discovery and identification of novel viroid sequences.

23.6 Conclusion

Viroids have been associated with major plant diseases and are emerging with expanding host range breadth. NGS has helped in easy detection and discovery of new viroid species. Life Sciences 454 high-throughput platform using sequence homology helped in the detection of many viroids along with their circularity confirmation using PAGE. Sequence homology tools such as VirFind, Virtool, and VirusDetect were used for detection from reads obtained. Different algorithms such as PFOR and PFOR2 were used for the detection of new viroids. The main function of the algorithm is to filter the overlapping regions from RNAs after deep sequencing from RNA pool for full-length sequence of viroid genome.

NGS has also been very useful in exploring the transcriptomic analyses of plants infected with viroids. This allowed in revelation of plant physiological processes affected due to viroid infection. The down- and upregulation of host genes can be used as targets for generating knockout/genome edited transgenics that may serve as viroid-resistant varieties. Third-generation sequencing is very helpful in sequencing as primers and amplification part is not included. Nanopore sequencing technology provides low-cost platform and is helpful in both detection and quantification of viroids. With the use of advanced sequencing methods, our understanding of viroid RNA biology has deepened, and the viroid research has exponentially progressed. More advanced techniques must come in future, so that the detection and discovery can be easy and cheap, and the functions of different genes can also be explored.

References

- Adams, I., & Fox, A. (2016). *Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. Current research topics in plant virology* (pp. 323–335). Cham: Springer.
- Adkar-Purushothama, C. R., Bolduc, F., Bru, P., & Perreault, J. P. (2020). Insights into potato spindle tuber viroid quasi-species from infection to disease. *Frontiers in microbiology*, *11*, 1235.
- Adkar-Purushothama, C. R., Kasai, A., Sugawara, K., Yamamoto, H., Yamazaki, Y., He, Y. H., . . . Sano, T. (2015). RNAi mediated inhibition of viroid infection in transgenic plants expressing viroid-specific small RNAs derived from various functional domains. *Scientific Reports*, *5*(1), 1–13.
- Adkar-Purushothama, C. R., & Perreault, J. P. (2020). Impact of nucleic acid sequencing on viroid biology. *International Journal of Molecular Sciences*, *21*(15), 5532.
- Adkar-Purushothama, C. R., Sano, T., & Perreault, J. P. (2018). Viroid-derived small RNA induces early flowering in tomato plants by RNA silencing. *Molecular Plant Pathology*, *19*(11), 2446–2458.
- Allison, L. A., Simon, L. D., & Maliga, P. (1996). Deletion of *rpoB* reveals a second distinct transcription system in plastids of higher plants. *The EMBO Journal*, *15*(11), 2802–2809.
- Antignus, Y., Lachman, O., Pearlsman, M., Gofman, R., & Bar-Joseph, M. (2002). A new disease of greenhouse tomatoes in Israel caused by a distinct strain of Tomato apical stunt viroid (TASVd). *Phytoparasitica*, *30*(5), 502–510.
- Barba, M., & Hadidi, A. (2017). *Application of next-generation sequencing technologies to viroids. Viroids and satellites* (pp. 401–412). Academic Press.
- Botermans, M., Roenhorst, J. W., Hooftman, M., Verhoeven, J. T. J., Metz, E., van Veen, E. J., & Westenberg, M. (2020). Development and validation of a real-time RT-PCR test for screening pepper and tomato seed lots for the presence of pospiviroids. *PLoS One*, *15*(9), e0232502.
- Brass, J. R., Owens, R. A., Matoušek, J., & Steger, G. (2017). Viroid quasispecies revealed by deep sequencing. *RNA Biology*, *14*(3), 317–325.
- de Noronha Fonseca, M. E., Marcellino, L. H., & Gander, E. (1996). A rapid and sensitive dot-blot hybridization assay for the detection of citrus exocortis viroid in *Citrus medica* with digoxigenin-labelled RNA probes. *Journal of Virological Methods*, *57*(2), 203–207.
- Di Serio, F., Malfitano, M., Flores, R., & Randles, J. W. (1999). Detection of peach latent mosaic viroid in Australia. *Australasian Plant Pathology*, *28*(1), 80–81.

- Fadda, Z., Daròs, J. A., Fagoaga, C., Flores, R., & Durán-Vila, N. (2003). Eggplant latent viroid, the candidate type species for a new genus within the family Avsunviroidae (hammerhead viroids). *Journal of Virology*, 77(11), 6528–6532.
- Flores, R., Hernández, C., Alba, A. E. M. D., Daròs, J. A., & Serio, F. D. (2005). Viroids and viroid-host interactions. *Annual Review of Phytopathology*, 43, 117–139.
- Flores, R., Hernandez, C., Llacer, G., & Desvignes, J. C. (1991). Identification of a new viroid as the putative causal agent of pear blister canker disease. *Journal of General Virology*, 72(6), 1199–1204.
- Flores, R., Navarro, B., Delgado, S., Serra, P., & Di Serio, F. (2020). Viroid pathogenesis: A critical appraisal of the role of RNA silencing in triggering the initial molecular lesion. *FEMS Microbiology Reviews*, 44(3), 386–398.
- Glouzon, J. P. S., Bolduc, F., Wang, S., Najmanovich, R. J., & Perreault, J. P. (2014). Deep-sequencing of the peach latent mosaic viroid reveals new aspects of population heterogeneity. *PLoS One*, 9(1), e87297.
- Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *The Journal of Investigative Dermatology*, 133(8), e11.
- Gucek, T., Trdan, S., Jakse, J., Javornik, B., Matousek, J., & Radisek, S. (2017). Diagnostic techniques for viroids. *Plant Pathology*, 66(3), 339–358.
- Hadidi, A. (2019). Next-generation sequencing and CRISPR/Cas13 editing in viroid research and molecular diagnostics. *Viruses*, 11(2), 120.
- Hadidi, A., & Yang, X. (1990). Detection of pome fruit viroids by enzymatic cDNA amplification. *Journal of Virological Methods*, 30(3), 261–269.
- Hammond, R., Smith, D. R., & Diener, T. O. (1989). Nucleotide sequence and proposed secondary structure of Columnnea latent viroid: a natural mosaic of viroid sequences. *Nucleic Acids Research*, 17(23), 10083–10094.
- Hosokawa, M., Matsushita, Y., Uchida, H., & Yazawa, S. (2006). Direct RT-PCR method for detecting two chrysanthemum viroids using minimal amounts of plant tissue. *Journal of Virological Methods*, 131(1), 28–33.
- Hou, W. Y., Li, S. F., Wu, Z. J., Jiang, D. M., & Sano, T. (2009). Coleus blumei viroid 6: A new tentative member of the genus Coleviroid derived from natural genome shuffling. *Archives of Virology*, 154(6), 993–997.
- Jakše, J., Radišek, S., Pokorn, T., Matoušek, J., & Javornik, B. (2015). Deep-sequencing revealed a CBCVd viroid as a highly aggressive pathogen on hop. *Plant Pathology*.
- Jones, S., Baizan-Edge, A., MacFarlane, S., & Torrance, L. (2017). Viral diagnostics in plants using next generation sequencing: Computational analysis in practice. *Frontiers in Plant Science*, 8, 1770.
- Kamble, A., & Khairkar, R. (2016). Basics of bioinformatics in biological research. *International Journal of Applied Sciences and Biotechnology*, 4(4), 425–429.
- Koganezawa, H. (1985). Transmission to apple seedlings of a low molecular weight RNA extracted from apple scar skin diseased trees. *Japanese Journal of Phytopathology*, 51(2), 176–182.
- Koltunow, A. M., Krake, L. R., Johnson, S. D., & Rezaian, M. A. (1989). Two related viroids cause grapevine yellow speckle disease independently. *Journal of General Virology*, 70(12), 3411–3419.
- Kulski, J. K. (2016). Next-generation sequencing—An overview of the history, tools, and “omic” applications. *Next generation sequencing—advances, applications and challenges*, 3–60.
- López-Carrasco, A., Ballesteros, C., Sentandreu, V., Delgado, S., Gago-Zachert, S., Flores, R., & Sanjuán, R. (2017). Different rates of spontaneous mutation of chloroplastic and nuclear viroids as determined by high-fidelity ultra-deep sequencing. *PLoS Pathogens*, 13(9), e1006547.
- Malfitano, M., Barone, M., Duran-Vila, N., & Alioto, D. (2005). Indexing of viroids in citrus orchards of Campania, Southern Italy. *Journal of Plant Pathology*, 115–121.
- Martinez, G., Donaire, L., Llave, C., Pallas, V., & Gomez, G. (2010). High-throughput sequencing of Hop stunt viroid-derived small RNAs from cucumber leaves and phloem. *Molecular Plant Pathology*, 11(3), 347–359.
- Martínez-Soriano, J. P., Galindo-Alonso, J., Maroon, C. J., Yucel, I., Smith, D. R., & Diener, T. O. (1996). Mexican papita viroid: Putative ancestor of crop viroids. *Proceedings of the National Academy of Sciences*, 93(18), 9397–9401.
- Matoušek, J., & Patzak, J. (2000). A low transmissibility of hop latent viroid through a generative phase of *Humulus lupulus* L. *Biologia Plantarum*, 43(1), 145–148.
- Matsushita, Y., Kanda, A., Usugi, T., & Tsuda, S. (2008). First report of a Tomato chlorotic dwarf viroid disease on tomato plants in Japan. *Journal of General Plant Pathology*, 74(2), 182–184.
- Matsuura, S., Matsushita, Y., Kozuka, R., Shimizu, S., & Tsuda, S. (2010). Transmission of Tomato chlorotic dwarf viroid by bumblebees (*Bombus ignitus*) in tomato plants. *European Journal of Plant Pathology*, 126(1), 111–115.
- Mehmood, M. A., Sehar, U., & Ahmad, N. (2014). Use of bioinformatics tools in different spheres of life sciences. *Journal of Data Mining in Genomics & Proteomics*, 5(2), 1.
- Mishra, A. K., Kumar, A., Mishra, D., Nath, V. S., Jakše, J., Kocábek, T., ... Matoušek, J. (2018). Genome-wide transcriptomic analysis reveals insights into the response to citrus bark cracking viroid (CBCVd) in hop (*Humulus lupulus* L.). *Viruses*, 10(10), 570.
- Moreno, M., Vázquez, L., López-Carrasco, A., Martín-Gago, J. A., Flores, R., & Briones, C. (2019). Direct visualization of the native structure of viroid RNAs at single-molecule resolution by atomic force microscopy. *RNA Biology*, 16(3), 295–308.
- Mumford, R. A., Walsh, K., & Boonham, N. (2000). A comparison of molecular methods for the routine detection of viroids. *EPPO Bulletin*, 30(3-4), 431–435.
- Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the basics of NGS: From mechanism to variant calling. *Current genetic medicine reports*, 3(4), 158–165.
- Owens, R. A., & Diener, T. O. (1981). Sensitive and rapid diagnosis of potato spindle tuber viroid disease by nucleic acid hybridization. *Science (New York, N.Y.)*, 213(4508), 670–672.

- Owens, R. A., Sano, T., & Duran-Vila, N. (2012). *Plant viroids: Isolation, characterization/detection, and analysis. Antiviral resistance in plants* (pp. 253–271). Totowa, NJ: Humana Press.
- Owens, R. A., Tech, K. B., Shao, J. Y., Sano, T., & Baker, C. J. (2012). Global analysis of tomato gene expression during Potato spindle tuber viroid infection reveals a complex array of changes affecting hormone signaling. *Molecular Plant-Microbe Interactions*, 25(4), 582–598.
- Qi, Y., & Ding, B. (2002). Replication of Potato spindle tuber viroid in cultured cells of tobacco and *Nicotiana benthamiana*: The role of specific nucleotides in determining replication levels for host adaptation. *Virology*, 302(2), 445–456.
- Qi, Y., & Ding, B. (2003). Inhibition of cell growth and shoot development by a specific nucleotide sequence in a noncoding viroid RNA. *The Plant Cell*, 15(6), 1360–1374.
- Ramesh, S. V., Yogindran, S., Gnanasekaran, P., Chakraborty, S., Winter, S., & Pappu, H. R. (2020). Virus and viroid-derived small RNAs as modulators of host gene expression: molecular insights into pathogenesis. *Frontiers in Microbiology*, 11.
- Sasaki, M., & Shikata, E. (1977). On some properties of hop stunt disease agent, a viroid. *Proceedings of the Japan Academy, Series B*, 53(3), 109–112.
- Schnell, R. J., Kuhn, D. N., Ronning, C. M., & Harkins, D. (1997). Application of RT-PCR for indexing avocado sunblotch viroid. *Plant Disease*, 81(9), 1023–1026.
- Serra, P., Eiras, M., Bani-Hashemian, S. M., Murcia, N., Kitajima, E. W., Daròs, J. A., . . . Duran-Vila, N. (2008). Citrus viroid V: Occurrence, host range, diagnosis, and identification of new variants. *Phytopathology*, 98(11), 1199–1204.
- Shamloul, A. M., Faggioli, F., Keith, J. M., & Hadidi, A. (2002). A novel multiplex RT-PCR probe capture hybridization (RT-PCR-ELISA) for simultaneous detection of six viroids in four genera: Apscaviroid, Hostuviroid, Pelamoviroid, and Pospiviroid. *Journal of Virological Methods*, 105(1), 115–121.
- Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, 122(1), e59.
- Štajner, N., Radišek, S., Mishra, A. K., Nath, V. S., Matoušek, J., & Jakše, J. (2019). Evaluation of disease severity and global transcriptome response induced by Citrus bark cracking viroid, Hop latent viroid, and their co-infection in hop (*Humulus lupulus* L.). *International journal of molecular sciences*, 20(13), 3154.
- St-Pierre, P., Hassen, I. F., Thompson, D., & Perreault, J. P. (2009). Characterization of the siRNAs associated with peach latent mosaic viroid infection. *Virology*, 383(2), 178–182.
- Suzuki, T., Fujibayashi, M., Hataya, T., Taneda, A., He, Y. H., Tsushima, T., . . . Sano, T. (2017). Characterization of host-dependent mutations of apple fruit crinkle viroid replicating in newly identified experimental hosts suggests maintenance of stem-loop structures in the left-hand half of the molecule is important for replication. *Journal of General Virology*, 98(3), 506–516.
- Teruo, S. A. N. O., Kobayashi, T., Ishiguro, A., & Motomura, Y. (2000). Two types of grapevine yellow speckle viroid 1 isolated from commercial grapevine had the nucleotide sequence of yellow speckle symptom-inducing type. *Journal of General Plant Pathology*, 66(1), 68–70.
- Tsushima, T., & Sano, T. (2018). A point-mutation of *Coleus blumei* viroid 1 switches the potential to transmit through seed. *Journal of General Virology*, 99(3), 393–401.
- Vadmalai, G., Hanold, D., Rezaian, M. A., & Randles, J. W. (2006). Variants of Coconut cadang-cadang viroid isolated from an African oil palm (*Elaeis guineensis* Jacq.) in Malaysia. *Archives of Virology*, 151(7), 1447–1456.
- Verhoeven, J. T., Roenhorst, J. W., & Owens, R. A. (2011). Mexican papita viroid and tomato planta macho viroid belongs to a single species in the genus Pospiviroid. *Archives of Virology*, 156, 1433–1437.
- Visvader, J. E., & Symons, R. H. (1983). Comparative sequence and structure of different isolates of citrus exocortis viroid. *Virology*, 130(1), 232–237.
- Visvader, J. E., & Symons, R. H. (1985). Eleven new sequence variants of citrus exocortis viroid and the correlation of sequence with pathogenicity. *Nucleic Acids Research*, 13(8), 2907–2920.
- Visvader, J. E., Forster, A. C., & Symons, R. H. (1985). Infectivity and in vitro mutagenesis of monomeric cDNA clones of citrus exocortis viroid indicates the site of processing of viroid precursors. *Nucleic Acids Research*, 13(16), 5843–5856.
- Walia, Y., Dhir, S., Bhadoria, S., Hallan, V., & Zaidi, A. A. (2012). Molecular characterization of Apple scar skin viroid from Himalayan wild cherry. *Forest Pathology*, 42(1), 84–87.
- Walia, Y., Dhir, S., Ram, R., Zaidi, A. A., & Hallan, V. (2014). Identification of the herbaceous host range of Apple scar skin viroid and analysis of its progeny variants. *Plant pathology*, 63(3), 684–690.
- Walia, Y., Dhir, S., Zaidi, A. A., & Hallan, V. (2015). Apple scar skin viroid naked RNA is actively transmitted by the whitefly *Trialeurodes vaporariorum*. *RNA Biology*, 12(10), 1131–1138.
- Wang, Y. (2021). Current view and perspectives in viroid replication. *Current Opinion in Virology*, 47, 32–37.
- Wu, Q., Ding, S. W., Zhang, Y., & Zhu, S. (2015). Identification of viruses and viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. *Annual Review of Phytopathology*, 53, 425–444.
- Zhang, C., Wu, Z., Li, Y., & Wu, J. (2015). Biogenesis, function, and applications of virus-derived small RNAs in plants. *Frontiers in microbiology*, 6, 1237.
- Zhang, Z., Qi, S., Tang, N., Zhang, X., Chen, S., Zhu, P., . . . Wu, Q. (2014). Discovery of replicating circular RNAs by RNA-seq and computational algorithms. *PLoS Pathogens*, 10(12), e1004553.
- Zhong, X., Archual, A. J., Amin, A. A., & Ding, B. (2008). A genomic map of viroid RNA motifs critical for replication and systemic trafficking. *The Plant Cell*, 20(1), 35–47.

Computational analysis for plant virus analysis using next-generation sequencing

Chitra Nehra¹, Rakesh Kumar Verma¹, Nikolay Manchev Petrov², Mariya Ivanova Stoyanova³, Pradeep Sharma⁴ and Rajarshi Kumar Gaur⁵

¹Department of Biosciences, Mody University of Science and Technology, Sikar, Rajasthan, India, ²Department of Natural Sciences, New Bulgarian University, Sofia, Bulgaria, ³Department of Plant Protection, Institute of Soil Science, Agrotechnologies and Plant Protection (ISSAPP) “Nikola Pushkarov”, Sofia, Bulgaria, ⁴ICAR- Indian Institute of Wheat & Barley Research, Karnal, Haryana, India, ⁵Department of Biotechnology, Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India

24.1 Introduction

Plant viruses are causing problems for food security by affecting crop quality and quantity across the world (Loebenstein, 2008; Soliman, Mourits, Oude, Lansink, & van der Werf, 2012). So to maintain plant protection and to attain food security, reliable methods for plant virus diagnosis are required. Plant virus detection methods are roughly divided into two categories: one is specific methods that are generally directed to one or more specific virus species like serological Enzyme-linked Immunosorbent Assay (ELISA) or molecular tests (Polymerase chain reaction (PCR)) and second is nonspecific methods like indicator test plants, electron microscopy. Specific methodologies require prior information of the pathogens while non-specific methods do not need prior details about virus being diagnosed. But these approaches only categorize viruses at genus level on the basis of the physical and biological properties revealed by viruses.

After the introduction of next-generation sequencing (NGS), detection of new viruses and host has increased significantly. NGS-based approaches provide a nonspecific method for virus detection without requirement of prior knowledge about test pathogens but give a specific result about species/strain (Adams & Fox, 2016).

In 2009 NGS was first used in discovery of novel DNA/RNA viruses and viroids (Adams et al., 2009; Al Rwahnih, Daubert, Golino, & Rowhani, 2015; Kreuze, 2014). In the same year, to investigate the role of RNAi in plant–viroid interactions (Navarro et al., 2009), and to study the pathogenesis of viroid-derived small RNAs (vd-sRNAs) from a chloroplast-replicating viroid (Di Serio et al., 2009), sequencing of vd-sRNAs were done using NGS. Since then it has been used in different plant virology studies comprising viral genome sequencing, analysis of plant viral diversity and evolution, ecological and epidemiological studies, detection and diagnosis of known/unknown viruses in host plants.

In NGS a large number of viroid small Ribonucleic acid (vsRNA) or vd-sRNA sequencing can be performed in a single run. These sequences can be reassembled to find out the nucleotide sequence of virus/viroid genome(s). On the other hand, these sequences can be used in comparison with the host genome to recognize genes that may be suppressed upon virus infection because of their local homology with the virus. Similarly, homology of viral satellite RNAs with healthy plants sRNAs allows to predict a possible consequence of their evolution from the host plant genome, which ultimately provides an idea about the origin of these pathogens (Zahid et al., 2015). NGS technologies are helpful in monitoring the emergence and spread of pathogens by genotyping of previously known or unknown viral isolates and setting the control measures based on information (Mahuku et al., 2015).

24.2 Development of next-generation sequencing technology

The first rapid DNA sequencing method was developed by Frederick Sanger using primer extension approach and published as “DNA sequencing with chain-terminating inhibitors” in 1977 (Sanger, Nicklen, & Coulson, 1977). Another DNA

sequencing method based on chemical degradation was also developed by Gilbert and Maxam (Maxam & Gilbert, 1977). These sequencing methods were called “first-generation” sequencing technologies. In early 21st century, new technologies were developed to overcome Sanger sequencing method limitations that allowed whole genome to be sequenced in one run by implementing steps, including fragmentation of genome into small sequences, random sampling of these small sequences, sequencing, and de novo assembly. These technologies are called second-generation or NGS technologies. In recent years, single molecular sequencer like PacBio and Nanopore sequencers has been developed, which are jointly called “third-/or fourth-generation” sequencing methods. These technologies are able to read <100-kb length sequences but have issues in sequencing fidelity and still under development (Suzuki, 2020). So, currently NGS technology is being most used in various basic and applied researches, including plant pathology and plant virology studies. By using NGS a large volume of data can be produced, and it delivers fast, cheaper, and precise results.

In 2000 first NGS technology was launched by Massively Parallel Signature Sequencing (MPSS) Lynx Therapeutics (the United States) Company which was later taken over by Illumina. In the case of MPSS, high-throughput data, including large amount of short DNA sequences, were basically used for cDNA sequencing to check the expression levels of different genes (Brenner et al., 2000).

In 2004 454 Life Sciences (Branford, CT, the United States) launched a new generation of sequencing technologies. This company made a sequencing machine that reduced the cost of sequencing sixfold comparing to automated Sanger sequencing.

In 2005 Solexa, which was in 2007 purchased by Illumina, marketed the sequencing by synthesis–based Genome Analyzer. This analyzer was based on reversible dye terminator technology and engineered polymerase (Bentley et al., 2008). The latest model of GAIIx can generate 85 billion bases of usable data in single run.

In 2005–06 Life Sciences that were purchased by Roche company (Basel, Switzerland) launched the 454 GS 20 Roche sequencing platform that could produce 20 million bases per run.

The 454 Life Sciences that was later taken over by Roche Company (with headquarter in Basel, Switzerland) developed a parallel version of pyrosequencing. In pyrosequencing, luciferase generates light on the addition of individual nucleotide to nascent DNA, this light is detected and resultant data generate the sequence reads (Margulies et al., 2005). Pyrosequencing gives intermediate read lengths and lesser price per base in comparison to other sequencing methods like Sanger sequencing, Illumina, and SOLiD (Schuster, 2008).

In 2005–06 the 454 GS 20 Roche sequencing platform was launched, which transformed the sequencing technologies as it was able to generate 20 million bases (20 Mbp). New model 454 GS – FLX + Titanium sequencing platform can generate 600 Mbp of sequence data per run with read lengths of up to 1000 bp. Roche launched small-sized GS junior sequencing platform system that is capable of producing 400-bp long sequencing reads in a quick run (Life Sciences, a Roche Company).

In later years, Illumina has launched HiSeq platform series, HiSeq 2500, HiSeq 2000, HiSeq 1500, and HiSeq 1000, which differ in run time, output, cluster generation, and maximum read lengths. Sequencing platform HiSeq 2500 is developed aiming high-throughput applications and can sequence a human genome in a day. HiSeq 2000 and HiSeq 2500 can generate 600 billion bases per run. In 2011 Illumina launched a benchtop platform MiSeq which is capable of generating 1.5 Gbp per run. NovaSeq, the latest high-output sequencing platform of Illumina, can generate 13 billion reads per run.

In 2011 Pacific Biosciences’ (Menlo Park, CA, the United States) single-molecule real-time (SMRT) sequencer and Life Technologies’ Ion Torrent sequencer were launched [PacBio (Pacific Bioscience)]. In the year 2012–13 Oxford Technologies’ Nanopore (Oxford, United Kingdom) single-molecule sequencer was released which has the ability to read ultralong single-molecule reads [Nanopore (Oxford Technologies)]. SMRT sequencer is capable of reading maximum read length >100,000 bases with 87% raw read accuracy (GenomeWeb, 2012) but it is quite expensive. SMRT sequencing techniques are also called “third-generation” or “long-read” sequencing.

Helicos BioSciences developed a single-molecule sequencing method that gives short ~35 bp reads (SeqLL, 2014) and it sequences nonamplified DNA, so to escape errors associated with amplification step (Heather & Benjamin, 2016). Other methods for DNA sequencing like microfluidic systems were also developed which can be used in RNA sequencing too (Zilionis et al., 2017). This indicates that many of these DNA sequencing methods can be useful tools for detail study of genomes and transcriptomes in future.

Additional methods of sequencing like Ion semiconductor sequencing, combinatorial probe anchor synthesis (BGI/MGI), Nanopore sequencing were also developed recently. Sequence read length of these methods are up to 600, 300, 2,272,580 bp, respectively (Loose, Rakyant, Holmes, & Payne, 2018; Fang et al., 2018).

In recent years, companies like Illumina, Qiagen, and Thermo Fisher Scientific are actively working in the development of high-throughput sequencing products (Straiton, Free, Sawyer, & Martin, 2019). In recent time, Illumina is

TABLE 24.1 Main features and performances of various next-generation sequencing (NGS) platforms.

NGS platform—company	Template preparation	Read length per run	Max output per run	Run time	No of reads per run	Chemistry	Raw error rate (%)
<i>First generation</i>							
Sanger—Life technologies	Primer extension	800 bp	84Kb	2 h	1	Dideoxy terminator	0.3
<i>Second generation</i>							
454GSFLX—Roche	Clonal-emPCR	700 bp	0.7 Gb	1–2 days	1×10^6	Pyrosequencing	1
GS Junior—Roche	Clonal-emPCR	500 bp	70 Mb	18 h	1×10^5	Pyrosequencing	
HiSeq—Illumina	Clonal bridge amplification	2×150	1500 Gb	0.3–11 days	5×10^9	Reversible dye terminators	0.8
MiSeq—Illumina	Clonal bridge amplification	2×300	15 Gb	27 h	3×10^8	Reversible dye terminators	0.8
SOLiD—Life technologies	Clonal-emPCR	50	120 Gb	14 days	1×10^9	Ligation	0.01
Retrovocity—BGI	Gridded DNA-nanoballs	50	3000 Gb	14 days	1×10^9	Hybridization/Ligation	0.01
Ion Proton—Life technologies	Clonal-emPCR	200	100 Gb	2–5 days	6×10^7	Proton detection	1.7
Ion PGM—Life technologies	Clonal-emPCR	200	2 Gb	2–5 h	5×10^6	Proton detection	1.7
<i>Third/or fourth generation</i>							
SMRT—Pacific Biosciences	Single molecule	>10,000	1 Gb	1–2 h	1×10^6	Real-time single-molecule sequencing	12.9
Helioscope—Helicos	Single molecule	35	25 Gb	8 days	7×10^9	Real-time single-molecule Sequencing	0.2
Nanopore—Oxford Nanopore technologies	Single molecule	>5000	1 Gb	2–3 days	6×10^4	Real-time single-molecule sequencing	34

considered to be the most popular platform for NGS technologies as it provides sequencing at affordable cost to diagnostic laboratories. Details about salient features and performance of main sequencing platforms are mentioned in Table 24.1 (Kwong, McCallum, Sintchenko, & Howden, 2015; Lam, Clark, & Chen, 2012; Li, Tighe, & Nicolet, 2014; Liu, Li, & Li, 2012; Metzker, 2010).

24.3 Next-generation sequencing data analysis by bioinformatics tools

Bioinformatics is a major limiting step for NGS technologies considering to overcome the growing challenges of storage, analysis, and interpretation of NGS data (Land, Hauser, & Jun, 2015). NGS data analysis software can be categorized into four general categories that are generation of sequence reads, base calling and/or polymorphism detection, de novo assembly of genome, and annotation. For each category, various software programs have been

developed. For example, to remove low quality and contaminant reads, an open-source software NGS QC Toolkit was developed by Patel and Jain that allows parallel sequencing of large amount of data (Patel & Jain, 2012). The resulting high-quality sequence reads are joined using de novo assembly bioinformatics tools to discover and reconstruction of novel genomes (Van der Walt, van Goethem, & Ramond, 2017). Various bioinformatics software and their practical applications are available for de novo assembly of short reads into whole genomes and transcriptomes (Paszkiwicz & Studholme, 2010; Shendure & Ji, 2008; Zhang, Chen, & Yang, 2011). Advanced and improved software are being commercially developed worldwide to analyze, assembly, and annotation of short reads. Detailed reviews about main features of available software's tools for quality control, genome assembly, taxonomic and functional annotation of data produced by NGS have been published in different articles (Almeida & De Martinis, 2019; Horner, Pavesi, & Castrignano, 2009; Miller, Koren, & Sutton, 2010; Pabinger, Dander, & Fischer, 2013). The tools and algorithms for NGS data analysis are continuously being developed and upgraded to keep pace with the latest advancement in these sequencing technologies.

24.4 Next-generation sequencing in plant virology

NGS and advanced bioinformatics tools have considerably added a number of new plant viruses/viroids detected and identified in host plants and vectors. In plant virology, NGS is being used in whole-genome sequencing, diversity and evolution of genome, ecology, epidemiology, transcription, replication, discovery, detection and identification. In recent years, hundreds of new DNA and RNA plant viruses belonging to different genera and families have been reported (Anja et al., 2017; Barba & Hadidi, 2015; Barba, Czosnek, & Hadidi, 2014; Gaafar & Ziebell, 2020; Hadidi & Barba, 2012; Ho & Tzanetakis, 2014; Roossinck, Martin, & Roumagnac, 2015; Wu et al., 2012). Since viroid RNA nature is noncoding, so a combination of methods, including biological indexing and molecular biology techniques, as well as plant certification and quarantine programs were being used in diagnosis of viroid (Gucek et al., 2017; Hadidi, Czosnek, & Barba, 2004; Owens, Sano, & Duran-Vila, 2012). NGS allows pathogen characterization without any prior knowledge about it thus assisting in detection and discovery of viroids (Barba et al., 2014; Li et al., 2012). One such example is the discovery of two new viroids, persimmon viroid and grapevine latent viroid, by using NGS (Ito, Suzaki, Nakano, & Sato, 2013; Zhang et al., 2014). In comparison with biological indexing method, NGS was found to be quick, sensitive, and extensive method for the detection of grapevine viruses (Al Rwahnih, Daubert, Golino, & Rowhani, 2009).

Plant viruses/viroids can be detected indirectly by sequencing small interfering RNA (siRNAs) in host plant. siRNAs are produced by host as defense mechanism in response to infection by viruses/viroids, which are small 21- to 24-nt RNA molecules that inactivate DNA/RNA viruses and viroids (Flores et al., 2015; Zhang, Wu, Li, & Wu, 2015). NGS reads of siRNAs provide information about pathogenic viruses/viroids that are previously unknown and present even at very low titers. A large number of vsRNA or vd-sRNA can be sequenced by NGS in a single run. To find out virus/viroids genome sequence, these small sequences can be reassembled. Sequencing of vsRNAs from nine different viruses from four different host plants provided detail information about distribution and composition of these vsRNAs in host plants and biogenesis of vsRNAs (Donaire et al., 2009). In the same year, two novel badnaviruses (double strand deoxyribose nucleic acid (dsDNA)) and one novel mastrevirus (single strand deoxyribose nucleic acid (ssDNA)) were identified (Kreuze et al., 2009). NGS offers to detect viruses that are not detectable by routine quarantine virus detection methods. One such example is discovery of sugarcane streak virus (genus *Mastrevirus*) from sugarcane plants which was escaped to be detected during routine quarantine virus detection methods (Candresse et al., 2014). Similarly, a new *Luteovirus* was discovered from introduced nectarine trees by using NGS (Bag et al., 2015; Villamor, Mekuria, & Eastwell, 2016). In Slovenia the reason behind severe stunting and death of hop plants was found to be citrus bark cracking viroid as revealed by NGS technologies (Jakse, Radisek, Pokorn, Matousek, & Javornik, 2015). These pathogens have been added in alert list for trading of plant material so that countries become aware of possible pathogen introduction. NGS could be thus helpful in controlling the introduction of foreign pathogens into a new country during the import of plant materials. Many novel and known viruses were detected from crop plants as well from wild hosts by using NGS in metagenomic methods (Roossinck et al., 2015; Roossinck, 2015; Stobbe & Roossinck, 2014). By using high-throughput sequencing techniques, complete genome sequencing of many known viruses were obtained which could be utilized for the identification and characterization of viral isolates in different novel and known host plants during infection. For example, complete genome sequencing of Artichoke latent virus by NGS identified that it is a member of genus *Macluravirus*, family Potyviridae, and ranunculus latent virus is not a different species but a strain of this virus (Minutillo et al., 2015); potato virus Y and S were detected in Maori potato (*Solanum tuberosum*) and turnip mosaic virus in rengarenga (*Arthropodium cirratum*) which were found to be a novel host (Blouin et al., 2016).

NGS is playing a major role in joining plant virology with other biological areas like CRISPER-Cas9-based genome-editing. CRISPER-Cas9 has been utilized in developing resistance against DNA and RNA viruses (Ali et al., 2015; Baltes et al., 2015; Ji, Zhang, Zhang, Wang, & Gao, 2015). An approach utilizing combination of NGS and CRISPER-Cas9 genome editing technique may help in controlling disease that is triggered by DNA/RNA viruses and viroids at the genomic level (Hadidi, Flores, Candresse, & Barba, 2016).

Conventional methods of virus detection in fruit trees are molecular, serological, and biological indexing, which are labor intensive and lengthy. NGS technologies are comparatively rapid and highly sensitive to detect disease-causing pathogens and potentially applicable in monitoring as well ensuring that trees are virus free. NGS was proved to be equally efficient in detecting known viral pathogens from fruit trees and was equal or superior in detecting novel viruses when compared to conventional viral detection methods (Rott et al., 2017). So, NGS-based approaches can be possibly replacing most or all conventional methods in terms of being quicker and more extensive than conventional methods with the same or more efficiency to detect fruit tree viruses. Applications of NGS in various studies of plant virus small RNAs, siRNAs, and viroid small RNAs are listed in Tables 24.2 and 24.3.

TABLE 24.2 List of application of next-generation sequencing (NGS) in various studies of plant virus small RNAs, siRNAs.

S. no.	Host	Results	Sample preparation	References
1.	Sweet potato	Identification of two novel badnaviruses (dsDNA) and one novel mastrevirus (ssDNA), detection of <i>Sweet potato feathery virus</i> and <i>sweet potato chlorotic stunt virus</i> .	siRNA	Kreuze et al. (2009)
2.	<i>Gomphrena globosa</i>	Novel <i>Gayfeather mild mottle virus</i> was discovered.	Total RNA	Adams et al. (2009)
3	<i>Arabidopsis thaliana</i>	<i>Tobacco mosaic virus</i> siRNA mediated virus–host interaction that may contribute to viral pathogenicity.	vsRNA	Qi et al. (2009)
4.	<i>Nicotiana benthamiana</i> , <i>A. thaliana</i> , <i>Cucumis melo</i> , and tomato	Nine different virus vsRNAs were studied in four different hosts. This study provided details about distribution and composition of vsRNA as well biogenesis of vsRNAs.	vsRNA	Donaire et al. (2009)
5.	Cassava	The complete genome sequence of the Tanzanian strain of <i>Cassava brown streak virus</i> was obtained.	Total RNA	Monger et al. (2010)
6.	Grapevine	<i>Grapevine syrah 1 virus</i> was discovered in grapevine and leafhopper vector.	Total RNA	Al Rwahnih et al. (2009)
7.	<i>N. benthamiana</i>	vsRNAs of <i>Cymbidium ringspot virus</i> were determined. These vsRNAs are primarily derived from positive strand of virus, had a 5' monophosphate, were accumulated with different frequencies, and not perfect duplexes.	vsRNA	Szittyta et al. (2010)
8.	<i>Oryza sativa</i>	Classification of vsRNAs from four rice stripe virus genome RNAs.	vsRNA	Yan et al. (2010)
9.	<i>N. benthamiana</i> , <i>A. thaliana</i>	Identification of vsRNAs and its associated satellite RNAs in <i>bamboo mosaic virus</i> .	vsRNA	Lin et al. (2010)
10.	Wild cocksfoot grass	<i>Cereal yellow dwarf virus (Luteovirus)</i> was discovered in wild cocksfoot grass and <i>Cocksfoot streak virus (Potyvirus)</i> was detected.	vsRNA	Pallett et al. (2010)
11.	Grapevine	Virus genera <i>Foveavirus</i> , <i>Maculavirus</i> , <i>Marafivirus</i> , and <i>Nepovirus</i> of vsRNAs were derived from both genomic and antigenomic strands, while genus <i>Tymovirus</i> of vsRNAs was originated from antigenomic strand.	vsRNA	Pantaleo et al. (2010)

(Continued)

TABLE 24.2 (Continued)

S. no.	Host	Results	Sample preparation	References
12.	Grapevine	Mycoviruses showing similarity with <i>Penicillium chrysogenum virus</i> , GLRaV-3, GRSPaV, GVA, and <i>Grapevine virus E</i> were detected.	dsRNA	Coetzee et al. (2010)
13.	Cotton	siRNA profiling of <i>Cotton leafroll dwarf virus (Poleovirus, Luteoviridae)</i> in infected cotton plants.	siRNA	Silva et al. (2011)
14.	Wild <i>Passiflora caerulea</i>	The complete genome sequence of <i>Passion fruit woodiness virus (Potyvirus)</i> was obtained.	Poly-A RNA	Wylie et al. (2011)
15.	Pepper, eggplant	The complete genome sequences of two novel viruses <i>Pepper yellow curl virus (Polerovirus)</i> and <i>Eggplant mild leaf mottle virus (Ipomovirus)</i> were obtained.	vsRNA	Dombrovsky et al. (2011)
16.	Tomato	Detection of <i>Tomato spotted wilt virus</i> in tomato at early infection period. Identification of <i>Tospovirus</i> and squash-infecting geminivirus; analysis of virus quasispecies.	siRNA	Hagen et al. (2011)
17.	<i>A. thaliana</i>	vsRNA and transcriptome profiling of <i>Arabidopsis</i> plants infected by <i>Oilseed rape mosaic virus</i> (genus <i>Tobamovirus</i>).	vsRNA	Hu et al. (2011)
18.	Tomato, <i>N. benthamiana</i>	Characterization of vsRNA of <i>Tomato yellow leaf curl virus</i> and associated betasatellite.	vsRNA	Yang et al. (2011)
19.	Citrus	vsRNA profiling reconstructed the full genome of T318A Spanish citrus tristeza virus isolate. vsRNAs map primarily at 3' end of genomic RNA.	vsRNA	Ruiz-Ruiz et al. (2011)
20.	<i>N. benthamiana</i> , <i>Laodelphax striatellus</i> (small brown leafhopper), rice	siRNAs of <i>Rice stripe virus</i> were found in infected rice, <i>Nicotiana</i> , and brown leafhopper.	siRNA	Xu et al. (2012)
21.	Tomato	Two strains of <i>Pepino mosaic virus</i> were identified and differentiated. Complete genome sequence of novel <i>Tomato necrotic stunt virus</i> was discovered.	vsRNA	Li et al. (2012)
22.	Sweet potato	Detection of different genera (<i>Potyvirus</i> , <i>Crinivirus</i> , <i>Begomovirus</i>). Analysis of vsRNA by NGS is a reliable and sensitive method for virus detection in crops.	vsRNA	Kashif et al. (2012)
23.	Citrus	Identification of <i>Citrus yellow vein clearing virus</i> in citrus (genus <i>Mandarivirus</i>).	siRNA	Loconsole et al. (2012a)
24.	Citrus	Identification of <i>Citrus chlorotic dwarf-associated virus</i> in citrus (genus <i>Begomovirus</i>)	siRNA and total DNA	Loconsole et al. (2012b)
25.	Apple	Two apricot viruses and four apple viruses associated with apple green crinkle were detected.	vsRNA	Yoshikawa et al. (2012)
26.	Grapevine	Detection of <i>Grapevine rupestris stem-pitting associated virus</i> , <i>Grapevine rupestris vein feathering virus</i> , and <i>Grapevine syrah virus</i> . A novel <i>Grapevine pinot gris virus</i> was discovered.	vsRNA	Giampetruzzi et al. (2012)

(Continued)

TABLE 24.2 (Continued)

S. no.	Host	Results	Sample preparation	References
27.	Grapevine	Characterization of vsRNAs associated with grapevine leafroll disease	vsRNA	Alebi et al. (2012)
28.	Apple, citrus, grapevine	ASPV, ACLSV, and an unknown mycovirus were detected. Two variants of CTV and ASGV were detected. Variants of GLRaV-3, GVA, and an unknown mycovirus were also detected.	siRNA	Maree et al. (2012)
29.	Cherry	Characterization of the genome of the divergent <i>Little cherry virus 1</i> (LChV1) isolate and establishing that LChV1 isolates could be responsible for Shirofugen stunt disease syndrome.	dsRNA	Candresse et al. (2014)
30.	Citrus	The complete nucleotide sequence of a novel virus <i>Citrus leprosis virus cytoplasmic type 2</i> (genus <i>Cilevirus</i>) was determined	siRNA	Roy et al. (2013)
31.	Citrus	The complete nucleotide sequence of novel <i>Citrus vein enation virus</i> was determined.	siRNA	Vives et al. (2013)
32.	Grapevine	Complete sequence of a novel single-stranded DNA virus <i>Grapevine red leaf-associated virus</i> was obtained.	Total RNA treated with DNase	Poojari et al. (2013)
33.	Black pepper	The complete genome sequence of <i>Piper yellow mosaic virus</i> (genus <i>Badnavirus</i> , family <i>Caulimoviridae</i> .) was determined. Partial sequences of two additional novel viruses <i>Piper DNA virus 1</i> and <i>2</i> were obtained.	Viral and plant DNAs were isolated from virus-enriched fraction	Hany et al. (2013)
34.	Potato (<i>Solanum tuberosum</i>)	<i>Potato virus Y</i> strains O, N, and NTN vsRNAs were different in same host that shows they interact differently. vsRNA were generated from every position in the genome.	vsRNA	Naveed et al. (2014)
35.	Potato (<i>S. tuberosum</i>)	Potato virus X siRNAs were separated according to their strains.	siRNA	Kutnjak et al. (2014)
35.	<i>A. thaliana</i> and <i>N. benthamiana</i>	Vs-RNAs and vd-sRNAs profiling allowed de novo reassembly of DNA and RNA viruses and viroids for the diagnosis and detection of known and novel virus.	vsRNA	Seguin et al. (2014)
36.	Apple	VsRNA profiling of apple stem grooving virus was done. The role of tRNA-derived sRNAs in plant–virus interaction was observed.	vsRNA	Visser et al. (2014)
37.	<i>Cucurbita pepo</i>	<i>Zucchini mosaic virus</i> vsRNAs were used to analyze the systemic movement of virus within inoculated leaf. With the distance from inoculation site, the number of virus variant increases.	vsRNA	Dunham et al. (2014)
38.	Sugarcane	Discovery of <i>Sugarcane streak virus</i> (genus <i>Mastrevirus</i>). The accumulation levels of vsRNAs are heavily influenced by both viral genomic ssDNA and its mRNA transcript secondary structure.	vsRNA	Candresse et al. (2014)
39.	Mulberry	Identification and molecular characterization of novel monopartite geminivirus associated with mulberry mosaic dwarf disease.	vsRNA	Ma et al. (2015)

(Continued)

TABLE 24.2 (Continued)

S. no.	Host	Results	Sample preparation	References
40.	Apple	Identification and molecular characterization of novel monopartite geminivirus.	vsRNA	Liang et al. (2015)
41.	Squash	Identification and molecular characterization of <i>Squash mosaic virus</i> .	vsRNA	Li et al. (2012)
42.	Chickpea (<i>Cicer arietinum</i>)	<i>Tomato mosaic virus</i> infection to chickpea in Europe	vsRNA	Pirovano et al. (2015)
43.	Melon, cucumber	Comparative vsRNAs analysis among source, sink, and phloem tissues in two different plant–virus pathosystems. Melon plants were infected with melon necrotic spot virus and cucumber plants were infected with prunus necrotic ringspot.	vsRNA	Herranz et al. (2015)
44.	Wild rose (<i>Rosa multiflora thumb.</i>)	Identification and molecular characterization of novel Closterovirus rose leaf rosette virus.	vsRNA	He et al., 2015
45.	Grapevine	The presence of eight different viruses was detected in one set of eight grapevines.	Small RNA	Eichmeier et al., 2016
46.	Tomato, mustard, potato, pea, tobacco	Twelve different viruses (<i>Potato virus Y</i> , <i>Cauliflower mosaic virus</i> , <i>Tomato yellow leaf curl virus</i> , <i>Alfalfa mosaic virus</i> , <i>pea necrotic yellow dwarf virus</i> , <i>Tobacco mosaic virus</i> , <i>Tomato chlorosis virus</i> , <i>Pepino mosaic virus</i> , <i>Potexvirus</i> , <i>Tomato mosaic virus</i> , etc.) were detected. A putative novel Cytorhabdovirus was discovered.	Small RNA, Ribosomal depleted total RNA	Pecman et al. (2017)
47.	Grapevine	Fifteen different viruses were detected and phylogenetic analysis showed diseases caused mainly because of infected propagating material.	vsRNA	Czotter et al. (2018)
48.	Peach	Six different virus genomes were obtained using transcriptomic data. Amount and copy number of viral RNA were also studied. Single-nucleotide variations in each viral genome were also analyzed.	vsRNA	Jo et al. (2018)
49.	Tomato, <i>Co. globosa</i>	Model plants were infected with virus and full viral genomes of <i>Pepino mosaic virus</i> and <i>Gayfeather mild mottle virus</i> were sequenced and identified.	cDNA	Adams et al. (2009)
50.	Peach	Eight different viruses were identified from a single palm tree. Peach virus D was reported first from China.	vsRNA	Xu et al. (2019)
51.	Tomato	Twenty-nine different viruses were identified from tomato plant with and without <i>Ty-1</i> gene. A gemycircularvirus (Genomoviridae), a new alpha-satellite, and two novel <i>Begomovirus</i> species were detected only from tomato without the <i>Ty-1</i> gene. A novel begomovirus was found exclusively in the <i>Ty-1</i> pool.	Viral ssDNA	de Nazaré Almeida dos Reis et al. (2020)

By NGS past disease epidemics and evolution of pathogens can be studied, using genetic material of plant viruses and viroids that has been isolated from dried plant samples of years old herbaria or museums. In a remarkable study, from an approximately 750-year-old barley grain, the whole-genome sequence of a barley stripe mosaic virus (BSMV)

TABLE 24.3 List of application of next-generation sequencing (NGS) in various studies of viroid small RNAs.

S. no.	Host	Findings of study	Sample	References
1.	Peach	Vd-sRNAs of <i>peach latent mosaic viroid</i> were used to analyze evolution and pathogenesis of viroid.	siRNA	Di Serio et al. (2009), Navarro et al. (2012a)
2.	Grapevine	Pathogenesis and plant–viroid interaction studies of <i>hop stunt viroid</i> , and <i>grapevine yellow speckle viroid</i> .	siRNA	Navarro et al. (2009)
3.	Grapevine	Identification of <i>Australian grapevine viroid</i> , <i>Hop stunt viroid</i> , and <i>Grapevine yellow speckle viroid</i> .	Total RNA or dsRNA	Al Rwahnih et al. (2009)
4.	<i>Nicotiana benthamiana</i>	RNA-dependent RNA polymerase 6 inhibits accumulation and prevents meristem invasion of <i>Potato spindle tuber viroid</i> which replicates in nuclei.	Plant and viroid siRNA	Di serio et al. (2010)
5.	Cucumber	Analysis of <i>Hop stunt viroid</i> pathway which involved in the biogenesis of the viroid siRNAs.	SiRNA	Martinez et al. (2010)
6.	Grapevine	Characterization of vd-sRNA of <i>hop stunt viroid</i> and <i>grapevine yellow speckle 2 viroid</i> .	siRNA	Alabi et al. (2012)
7.	Tomato	Detection and identification of Potato spindle tuber viroid.	siRNA	Li et al. (2012)
8.	Grapevine	Detection and identification of <i>Grapevine yellow speckle viroid 1</i> and <i>Hop stunt viroid</i> .	siRNA	Giampetruzzi et al. (2012)
9.	Grapevine	Detection and identification of <i>Grapevine yellow speckle viroid 1</i> and <i>Hop stunt viroid</i> .	dsRNAs and siRNAs	Chiumenti et al. (2012)
10.	Grapevine	It was found that viroid-infected plants generate 21- to 24-nt vd-sRNAs. Based upon this, an approach was developed for identification of previously known and unknown viroids.	siRNA	Wu et al. (2012)
11.	Grapevine	Detection and identification of <i>Grapevine yellow speckle viroid 1</i> , <i>Hop stunt viroid</i> , <i>Citrus exocortis Yucatan viroid</i> , and <i>Citrus exocortis viroid</i> .	Total RNA treated with Dnase	Poojari et al. (2013)
12.	Different hosts	Viroid circular RNAs and satellite sRNAs were identified using bioinformatics tools.	RNA seq.	Zhang et al. (2014)
13.	Fig (<i>Ficus carica</i>)	Detection of apple dimple fruit viroid in new host fig.	vd-sRNAs	Chiumenti et al. (2014)
14.	Peach	Upon inoculation with a single variant of <i>peach latent mosaic viroid</i> generates a highly heterogeneous progeny within a single infected peach seedling.	vd-sRNAs	Glouzon et al. (2014)
15.	Tomato	<i>Potato spindle tuber viroid</i> (PSTVd) vd-sRNAs and effect of artificial miRNAs that were generated from PSTVd—mild or severe infected plants were studied. Analysis of distribution of vd-sRNAs hot spot indicates vd-sRNAs involvement in symptom expression.	vd-sRNAs	Adkar-Purushothama et al. (2015), Avina-Padilla et al. (2015)
16.	Chickpea	Detection of <i>hop stunt viroid</i> in new host chickpea.	vd-sRNAs	Pirovano et al. (2015)
17.	Grapevine	<i>Grapevine yellow speckle viroid 1</i> and <i>Hop stunt viroid</i> were detected.	vd-sRNAs	Eichmeier et al. (2016)
18.	Tomato, <i>Prunus</i> species	Three viroid species, including <i>Columnea latent viroid</i> , <i>Peach latent mosaic viroid</i> , <i>Tomato apical stunt viroid</i> , were reported.	Small RNA, Ribosomal depleted total RNA	Pecman et al. (2017)
19.	Grapevine	Three viroids Hop stunt viroid (HSVd) and Grapevine yellow speckled viroid 1–2 (GYSVd-1 and 2) were detected.	vd-sRNA	Czotter et al. (2018)

(Continued)

TABLE 24.3 (Continued)

S. no.	Host	Findings of study	Sample	References
20.	Peach	Two different viroids <i>Hop stunt viroid</i> and <i>peach latent mosaic viroid</i> were identified using transcriptome data. Amount of viroid RNA and copy number were analyzed.	Small RNA	Sen Lian et al. (2018)
21.	Apple	<i>Apple chlorotic fruit spot viroid</i> (genus <i>Apscaviroid</i>) was discovered.	Total RNA	Leichtfried et al. (2019)
22.	Peach	<i>Peach latent mosaic viroid</i> (PLMVd), sequences were isolated from symptomatic and asymptomatic peach leaves.	vsRNA	Xu et al. (2019)

isolate was obtained using NGS of small RNA sequences (Smith et al., 2014). Surprisingly, obtained sequence shows position differently in phylogenetic tree of BSMV, suggesting recent origin of the virus isolate. Similarly, a viral genome was discovered from 700-year-old caribou feces from a subarctic ice patch (Ng et al., 2014) and a huge DNA virus named *Pithovirus sibericum* from 30,000-year-old Siberian permafrost sample was also detected and identified through NGS (Legendre et al., 2014). Such findings would provide an insight into evolution of plant viruses and viroids.

24.5 Challenges

NGS offers an efficient and fast DNA, or RNA high-throughput sequencing of the complete genomes of plant viral/viroid pathogens and of the certain small RNAs produced during the infection process. High-throughput sequencing sRNAs followed by bioinformatics analysis is a significant method to detect and identify known as well novel plant viruses or viroids. This has been proved to be a powerful tool in the area of plant virus/viroids discovery and diagnosis (Massart, Olmos, Jijakli, & Candresse, 2014). Sequencing small vRNAs by NGS methods has been a universal approach, applied in approximately half of the published studies on plant viral diagnosis and disease symptom studies (Barba et al., 2014). But still in many experiments, either partial or relatively short sequences are obtained which may be due to lack of sufficient proportions of viral sRNA concentration in total sRNA as a result of which viruses are not detected. During the assembly of viral sRNA sequences into contigs, assemblers face computational challenges like great diversity of infecting viral population in a sample or relatively very less size of viral sRNAs in a large-sized host sRNAs reads. The quality of preliminary sRNA assembly is very important for the effective detection of a novel pathogen that has not been yet submitted in reference sequence databases (Barrero et al., 2017). Although theoretically it seems very simple for diagnosis of a pathogen by sRNA dataset analysis, practically it is a complex research experiment (Soueidan, Schmitt, Candresse, & Nikolski, 2014; Wu, Ding, Zhang, & Zhu, 2015). The ability of a sensitive, accurate, and replicable diagnosis during sRNAs analysis depends on setting a general strategy to opting specific tool and parameters for experiment. The outcomes of viral small RNA analysis for discovery and diagnosis of virus depend on characteristics of the obtained virus sequences and their accuracy, extensiveness, the pipeline performance, and expertise of scientists (Massart et al., 2019). Bioinformatics analysis strategy should be selected considering various factors during experiment and relative concentration of viral sequence in sequence dataset directly correlates with the sensitivity of results.

In testing pathogen infection in fruit trees, NGS is equally efficient and more rapid than conventional testing methods with no false negatives, but still a false positive is some concern in NGS-based diagnostic approaches. NGS-based techniques are complementary only not exclusive, so bioassay should be followed after NGS to confirm the results. Moreover, NGS can detect contaminant pathogen sequences which may not be essentially replicating at the host plant cell from where it was isolated (Blawid, Silva, & Nagata, 2017).

To analyze NGS sequence reads, various bioinformatics tools are used to identify pathogens and symptoms etiology. To develop these tools for analysis of RNA sequence data after NGS, various issues like uploading bulky NGS raw reads, intensive data processing steps on computer, dependence upon already processed custom database, etc. are being faced by researchers. Other than this, to identify novel virus scientific expertise as well frequent assembly and mapping

of viral sequences are required. To make tools available for effective, accurate, and easy analysis, NGS raw dataset is a big challenge. To process data on locally installed computer programs, there is no need to upload data on a remote server but requires expertise in software installation and operation. The webserver can be the most convenient option to a nonexpert user but it needs advanced computer hardware at the remote site and large data files have to be uploaded to process (Jones, Amanda, Stuart, & Lesley, 2017).

Plant virology has indeed advanced using new technologies like NGS but, at the same time, faces challenges like biological characterization of novel viruses and evaluation of their effect on biosecurity, monitoring, and scientific levels. The biological characterization of novel virus may be difficult with multiplex viruses, where viruses can transform their pathogenic potential by mutualistic interactions (Syller & Grupa, 2016). With the more advancement in sequencing technologies and bioinformatics tools, downstream epidemiology and disease etiology analysis will be quite challenging in determining biological significance and impact of a new virus or mixture of different viruses (Massart et al., 2017).

24.6 Conclusion and future prospective

NGS combined with bioinformatics tools has been providing a rapid sensitive and extensive method for DNA or RNA sequencing of plant viruses/viroids. Whole-genome sequencing of viruses and viroids is helpful in discovery and diagnosis of pathogens as well diversity, evolution, ecology, and virus–host interaction analysis in plant virology. Many known and unknown plant viruses and viroids are detected, identified, and classified from cultivated as well wild hosts using NGS techniques in the last few years. Since NGS method does not require any prior knowledge of pathogens, it has become a universal method for discovery and diagnosis of pathogen diagnosis. NGS is a powerful diagnostic tool that offers a deep insight into virus infection and is helpful in monitoring pathogen infection in vegetative propagating tree plants and during import/export of plant materials. In coming years, NGS can be used in increasing capabilities and reliabilities of plant quarantine and certification programs. NGS has been showing promising role in developing virus resistance in plants using genome-editing techniques. A combination of NGS and genome-editing techniques like CRISPER-Cas9 system can be utilized in developing resistance against DNA/RNA viruses and viroids and controlling pathogen infection at the genomic level. Since viruses cause significant loss in crop yield and affect the quality of plant products, rapid diagnosis of pathogen by NGS could be helpful in successful crop production and combating negative economic impacts.

NGS allows unbiased and hypothesis-free simultaneous detection of multiple viruses in plant samples. It can detect viruses even from samples where multiple viral infection present or unclear and unspecified disease symptoms are present. NGS technologies may become a game changer in the field of plant virology with more advancement in efficient nucleic acid extraction protocols, robust bioinformatics tools, and with availability at affordable cost.

References

- Adams, I., & Fox, A. (2016). Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. In A. Wang, & X. Zhou Cham (Eds.), *Current research topics in plant virology* (pp. 323–335). Springer.
- Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., et al. (2009). Next-generation sequencing and metagenomic analysis: A universal diagnostic tool in plant virology. *Molecular Plant Pathology*, *10*, 537–545.
- Adkar-Purushothama, C. R., Brosseau, C., Giguère, T., Sano, T., Moffett, P., & Perreault, J.-P. (2015). Small RNA derived from the virulence modulating region of the potato spindle tuber viroid silences callose synthase genes of tomato plants. *Plant Cell*, *27*, 2178–2194.
- Alabi, O. A., Haruna, M. T., Anokwuru, C. P., Jegede, T., Abia, H., Okegbe, V., et al. (2012). Comparative studies on antimicrobial properties of extracts of fresh and dried leaves of *Carica papaya* (L) on clinical bacterial and fungal isolates. *Pelagia Research Library*, *3*(5), 3107–3114.
- Al Rwahnih, M., Daubert, S., Golino, D., & Rowhani, A. (2009). Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology*, *387*, 395–401.
- Al Rwahnih, M., Daubert, S., Golino, D., & Rowhani, A. (2015). Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Phytopathology*, *105*, 758–763.
- Ali, Z., Abulfaraj, A., Idris, A., Ali, S., Tashkandi, M., & Mahfouz, M. M. (2015). CRISPR/Cas9-mediated viral interference in plants. *Genome Biology*, *16*, 238.
- Almeida, O. G. G., & De Martinis, E. C. P. (2019). Bioinformatics tools to assess metagenomic data for applied microbiology. *Applied Microbiology and Biotechnology*, *103*, 69–82.
- Anja, P., Denis, K., Ion, G. A., Ian, A., Adrian, F., Neil, B., et al. (2017). Next Generation Sequencing for detection and discovery of plant viruses and viroids: Comparison of two approaches. *Frontiers in Microbiology*, *8*, 1998.

- Avina-Padilla, K., Martínez de la Vega, O., Rivera-Bustamante, R., Martínez-Soriano, J. P., Owens, R. A., Hammond, R. W., et al. (2015). In silico prediction and validation of potential gene targets for pospiviroid-derived small RNAs during tomato infection. *Gene*, *564*(2), 197–205.
- Bag, S., Al Rwahnih, M., Li, A., Gonzalez, A., Rowhani, A., Uyemoto, J. K., et al. (2015). Detection of a new Luteovirus in imported nectarine trees: A case study to propose adoption of metagenomics in post-entry quarantine. *Phytopathology*, *105*, 840–846.
- Baltes, N. J., Hummel, A. W., Konecna, E., Cegan, R., Bruns, A. N., Bisaro, D. M., et al. (2015). Conferring resistance to geminiviruses with the CRISPR-Cas prokaryotic immune system. *Nature Plants*, *1*, 15145.
- Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, *6*, 106–136.
- Barba, M., & Hadidi, A. (2015). An overview of plant pathology and application of next-generation sequencing technologies. *CAB Reviews*, *10*, 1–21.
- Barrera, R. A., Napier, K. R., Cunnington, J., Liefing, L., Keenan, S., Frampton, R. A., et al. (2017). An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. *BMC Bioinformatics*, *18*, 26.
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59.
- Blawid, R., Silva, J. M. F., & Nagata, T. (2017). Discovering and sequencing new plant viral genomes by next-generation sequencing: Description of a practical pipeline. *Annals of Applied Biology*, *170*, 301–314.
- Blouin, A. G., Ross, H. A., Hobson-Peters, J., O'Brien, C., Warren, B., & MacDiarmid, R. (2016). A new virus discovered by immunocapture of double-stranded RNA, a rapid method for virus enrichment in metagenomic studies. *Molecular Ecology Resources*, *16*, 1255–1263.
- Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, *18*(6), 630–634.
- Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., et al. (2014). Appearances can be deceptive: Revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS One*, *9*(7), e102945.
- Chiumenti, M., Torchetti, E. M., Di Serio, F., & Minafra, A. (2014). Identification and characterization of a viroid resembling apple dimple fruit viroid in fig (*Ficus carica* L.) by next generation sequencing of small RNAs. *Virus Research*, *188*, 54–59.
- Coetzee, B., Freeborough, M. J., Maree, H. J., Celton, J. M., Rees, D. J. G., & Burger, J. T. (2010). Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. *Virology*, *400*, 157–163.
- de Nazaré Almeida dos Reis, L., Fonseca, M. E. D. N., Ribeiro, S. G., Naito, F. Y. B., Boiteux, L. S., & Pereira-Carvalho, R. D. C. (2020). Metagenomics of Neotropical Single-Stranded DNA Viruses in Tomato Cultivars with and without the Ty-1 Gene. *Viruses*, *12*(8), 819.
- Di Serio, F., Gisel, A., Navarro, B., Delgado, S., Martínez de Alba, A. E., Donvito, G., et al. (2009). Deep sequencing of the small RNAs derived from two symptomatic variants of a chloroplastic viroid: Implications, for their genesis and for pathogenesis. *PLoS One*, *4*, e7539.
- Di Serio, F., Martínez de Alba, A. E., Navarro, B., Gisel, A., & Flores, R. (2010). RNA-dependent RNA polymerase 6 delay accumulation and precludes meristem invasion of a viroid that replicates in the nucleus. *Journal of Virology*, *84*(5), 2477–2489.
- Dombrovsky, A., Glanz, E., Sapkota, R., Lachman, O., Bronstein, M., Schnitzer, T., et al. (2011). Next-generation sequencing a rapid and reliable method to obtain sequence data of the genomes of un described sequence data of the genomes of undescribed plant viruses; Proceedings of the BARD-Sponsored Workshop—Microarrays and Next-Generation Sequencing for Detection and Identification of Plant Viruses; Beltsville, MD, USA. 17–19 November 2011; Abstract No. 24.
- Donaire, L., Wang, Y., González-Ibeas, D., Mayer, K. F., Aranda, M. A., & Llave, C. (2009). Deep sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. *Virology*, *392*, 203–214.
- Dunham, J. P., Simmons, H. E., Holmes, E. C., & Stephenson, A. G. (2014). Molecular analysis of viral (zucchini yellow mosaic virus) genetic diversity during systemic movement through a Cucurbita pepo vine. *Virus Research*, *191*, 172–179.
- Eichmeier, A., Penazova, E., Pavelkova, R., Mynarzova, Z., & Saldarelli, P. (2016). Detection of Grapevine Pinot gris virus in certified grapevine stocks in Moravia, Czech Republic. *Journal of Plant Pathology*, *98*, 155–157.
- Flores, R., Minoia, S., Carbonell, A., Gisel, A., Delgado, S., López-Carrasco, A., et al. (2015). Viroids, the simplest RNA replicons: How they manipulate their hosts for being propagated and how their hosts react for containing the infection. *Virus Research*, *209*, 136–145.
- Gaafar, Y. Z. A., & Ziebell, H. (2020). Comparative study on three viral enrichment approaches based on RNA extraction for plant virus/viroid detection using high-throughput sequencing. *PLoS One*, *15*(8), e0237951.
- Fang, C., Zhong, H., Lin, Y., Chen, B., Han, M., Ren, H., et al. (2018). Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. *Gigascience*, *7*(3), 1–8.
- GenomeWeb. (2012). *After a year of testing, two early PacBio customers expect more routine use of RS sequencer in 2012.*
- Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., et al. (2012). A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. *Virus Research*, *163*, 262–268.
- Glouzon, J. P. S., Bolduc, F., Wang, S., Najmanovich, R. J., & Perreault, J. P. (2014). Deep-sequencing of the peach latent mosaic viroid reveals new aspects of population heterogeneity. *PLoS One*, *9*, e87297.
- Gucek, T., Trdan, S., Jakse, J., Javornik, B., Matousek, J., & Radisek, S. (2017). Diagnostic techniques for viroids. *Plant Pathology*, *66*, 339–358.
- Hadidi, A., & Barba, M. (2012). Next-generation sequencing: Historical perspective and current applications in plant virology. *Petria*, *22*, 262–277.
- Hadidi, A., Czosnek, H., & Barba, M. (2004). DNA microarrays and their potential applications for the detection of plant viruses, viroids, and phytoplasmas. *Journal of Plant Pathology*, *97*–104.
- Hadidi, A., Flores, R., Candresse, T., & Barba, M. (2016). Next-generation sequencing and genome editing in plant virology. *Frontiers in Microbiology*, *7*, 1325.

- Hagen, C., Frizzi, A., Kao, J., Jia, L., Huang, M., Zhang, Y., et al. (2011). Using small RNA sequences to diagnose, sequence, and investigate the infectivity characteristics of vegetable-infecting viruses. *Archives of Virology*, *156*, 1209–1216.
- Hany, U., Adams, I. P., Glover, R., Bhat, A. I., & Boonham, N. (2013). The complete nucleotide sequence of Piper yellow mottle virus (PYMoV). *Archives of Virology*, 158.
- He, Y., Yang, Z., Hong, N., Wang, G., Ning, G., & Xu, W. (2015). Deep sequencing reveals a novel closterovirus associated with wild rose leaf rosette disease. *Molecular Plant Pathology*, *16*, 449–458.
- Heather, J. M., & Benjamin, C. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8.
- Herranz, M. C., Navarro, J. A., Sommen, E., & Pallás, V. (2015). Comparative analysis among the small RNA populations of source, sink and conductive tissues in two different plant-virus pathosystems. *BMC Genomics*, *16*, 117.
- Ho, T., & Tzanetakis, J. E. (2014). Development of a virus detection and discovery pipeline using next generation sequencing. *Virology*, *47*, 54–60.
- Horner, D. S., Pavesi, G., Castrignano, T., et al. (2009). Bioinformatics approaches for genomics and post genomics applications of next generation sequencing. *Briefings in Bioinformatics*, *11*, 181–197.
- Hu, Q., Hollunder, J., Niehl, A., Korner, C. J., Gereige, D., Windels, D., et al. (2011). Specific impact of tobamavirus infection on the Arabidopsis small RNA profile. *PLoS One*, *6*, e19549.
- Ito, T., Suzuki, K., Nakano, M., & Sato, A. (2013). Characterization of a new apscaviroid from American persimmon. *Archives of Virology*, *158*, 2629–2631.
- Jakse, J., Radisek, S., Pokorn, T., Matousek, J., & Javornik, B. (2015). Deep-sequencing revealed Citrus bark cracking viroid (CBCVd) as a highly aggressive pathogen on hop. *Plant Pathology*, *64*, 831–842.
- Ji, X., Zhang, H., Zhang, Y., Wang, Y., & Gao, C. (2015). Establishing a CRISPR-Cas-like immune system conferring DNA virus resistance in plants. *National Plants*, *1*, 15144.
- Jo, Y., Lian, S., Chu, H., et al. (2018). Peach RNA viromes in six different peach cultivars. *Scientific Reports*, *8*, 1844.
- Jones, S., Amanda, B. E., Stuart, M., & Lesley, T. (2017). Viral diagnostics in plants using next generation sequencing: Computational analysis in practice. *Frontiers in Plant Science*, *8*, 1770.
- Kashif, M., Pietila, S., Artola, K., Tugume, A. K., Makinen, V., & Valkonen, J. P. T. (2012). Detection of viruses in sweetpotato from Honduras and Guatemala augmented by deep-sequencing of small-RNAs. *Plant Disease*, *96*, 1430–1437.
- Kreuze, J. (2014). siRNA deep sequencing and assembly: Piecing together viral infections. In M. L. Gullino, & P. J. M. Bonants (Eds.), *Detection and diagnostics of plant pathogens* (pp. 21–38). Dordrecht: Springer.
- Kreuze, J. F., Pérez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., et al. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology*, *388*, 1–7.
- Kutnjak, D., Silvestre, R., Cuellar, W., Perez, W., Müller, G., Ravnkar, M., et al. (2014). Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach. *Virus Research*, *191*, 45–50.
- Kwong, J. C., McCallum, S., Sintchenko, V., & Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology*, *47*, 199–210.
- Lam, H. Y., Clark, M. J., Chen, R., et al. (2012). Performance comparison of whole-genome sequencing platform. *Nature Biotechnology*, *30*, 78–83.
- Land, M., Hauser, L., Jun, S. R., et al. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, *15*, 141–161.
- Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., et al. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 4274–4279.
- Leichtfried, T., Dobrovolny, S., Reizenzein, H., Steinkellner, S., & Gottsberger, R. A. (2019). Apple chlorotic fruit spot viroid: a putative new pathogenic viroid on apple characterized by next-generation sequencing. *Archives of Virology*, *164*(12), 3137–3140.
- Li, R., Gao, S., Hernandez, A. G., Wechter, W. P., Fei, Z., & Ling, K. S. (2012). Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. *PLoS One*, *7*, e37127.
- Li, S., Tighe, S. W., Nicolet, C. M., et al. (2014). Multi-platform and cross-methodological reproducibility of transcriptome profiling by RNA-seq in the ABRF next generation sequencing study. *Nature Biotechnology* *Nat Biotechnol*, *32*, 915–925.
- Liang, P., Navarro, B., Zhang, Z., Wang, H., Lu, M., Xiao, H., et al. (2015). Identification and characterization of a novel geminivirus with a monopartite genome infecting apple trees. *Journal of General Virology*, *96*, 2411–2420.
- Life Sciences, a Roche Company. Available from <http://www.454.com>.
- Lin, K. Y., Cheng, C. P., Chang, B. C. H., Wang, W. C., Huang, Y. W., Lee, Y. S., et al. (2010). Global analysis of small interfering RNAs derived from Bamboo mosaic virus and its associated satellite RNAs in different plants. *PLoS ONE*, *5*, e11928.
- Liu, L., Li, Y., Li, S., et al. (2012). Comparison of next-generation sequencing system. *Journal of Biomedicine & Biotechnology*, *2012*, 251364.
- Loconsole, G., Onelge, N., Potere, O., Giampetruzzi, A., Bozan, O., Satar, S., et al. (2012a). Identification and characterization of Citrus yellow vein clearing virus, a putative new member of the genus Mandarivirus. *Phytopathology*, *102*, 1168–1175.
- Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G. P., & Saponari, M. (2012b). Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member of the family Geminiviridae. *Virology*, *432*, 162–172.
- Loebenstein, G. (2008). Plant virus diseases: Economic aspects. In M. H. V. Van Regenmortel, & W. J. Mahy Brian (Eds.), *Desk encyclopedia of plant and fungal virology* (pp. 426–430). Oxford: Academic Press.
- Loose, M., Rakyán, V., Holmes, N., & Payne, A. (2018). Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*, *10*.

- Ma, Y., Navarro, B., Zhang, Z., Lu, M., Zhou, X., Chi, S., et al. (2015). Identification and molecular characterization of a novel monopartite gemini-virus associated with mulberry mosaic dwarf disease. *Journal of General Virology*, *96*, 2421–2434.
- Mahuku, G., Lockhart, B. E., Wanjala, B., Jones, M. W., Kimunye, J. N., & Stewat, L. R. (2015). Maize lethal necrosis (mln), an emerging threat to maize-based food security in sub-Saharan Africa. *Phytopathology*, *105*(7), 956–965.
- Maree, H.J., Nel, Y., Visser, M., Coetzec, B., Manicom, B., Burger, J.T., et al. The study of plant virus disease etiology using next-generation sequencing technologies; Proceedings of the 22nd International Conference on Virus and Other Transmissible Diseases of Fruit Crops; Rome, Italy. 3–8 June 2012; Abstract No. 48.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bembe, L. A., et al. (2005). Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, *437*(7057), 376–380.
- Martinez, G., Donaire, L., Llave, C., Pallas, V., & Gomez, G. (2010). High-throughput sequencing of Hop stunt viroid-derived small RNAs from cucumber leaves and phloem. *Molecular Plant Pathology*, *11*, 347–359.
- Massart, S., Chiumenti, M., Jonghe, K. D., Glover, R., Haegeman, A., Koloniuk, I., et al. (2019). Virus detection by high throughput sequencing of small RNAs: Large scale performance testing of sequence analysis strategies. *Phytopathology*, *109*, 488–497.
- Massart, S., Olmos, A., Jijakli, H., & Candresse, T. (2014). Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research*, *188*, 90–96.
- Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnika, M., et al. (2017). A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Frontiers in Microbiology*, *8*, 45.
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564.
- Metzker, M. L. (2010). Sequencing technologies-The next generation. *Nature Reviews Genetics*, *11*, 31–46.
- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next generation sequencing data. *Genetics*, *95*, 315–327.
- Minutillo, S. A., Marais, A., Mascia, T., Faure, C., Svanella-Dumas, L., Theil, S., et al. (2015). Complete nucleotide sequence of Artichoke latent virus shows it to be a member of the genus *Macluravirus* in the family Potyviridae. *Phytopathology*, *105*, 1155–1160.
- Monger, W. A., Alicai, T., Ndunguru, J., Kinyua, Z. M., Potts, M., Reeder, et al. (2010). The complete genome sequence of the Tanzanian strain of Cassava brown streak virus and comparison with the Ugandan strain sequence. *Archives of virology*, *155*(3), 429–433.
- Nanopore (Oxford Technologies). Available from <<https://www.nanoporetech.com/>> Accessed 12.11.20.
- Navarro, B., Pantaleo, V., Gisel, A., Moxon, S., Dalmay, T., Bistray, G., et al. (2009). Deep sequencing of viroid-derived small RNAs from grapevine provides new insight on the role of RNA silencing in plant-viroid interaction. *PLoS One*, *4*, e7686.
- Navarro, B., Gisel, A., Rodio, M. E., Delgado, S., Flores, R., & Di Serio, F. (2012a). Small RNAs containing the pathogenic determinant of a chloroplast-replicating viroid guide the degradation of a host mRNA as predicted by RNA silencing. *Plant Journal*, *70*, 991–1003.
- Naveed, K., Mitter, N., Harper, A., Dhingra, A., & Pappua, H. R. (2014). Comparative analysis of virus-specific small RNA profiles of three biologically distinct strains of potato virus Y in infected potato (*Solanum tuberosum*) cv. Russet Burbank. *Virus Research*, *191*, 153–160.
- Ng, T. F. F., Chen, L. F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P. D., et al. (2014). Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 16106–16111.
- Owens, R. A., Sano, T., & Duran-Vila, N. (2012). Plant viroids: Isolation, characterization/detection, and analysis. *Methods in Molecular Biology*, *894*, 253–271.
- Pabinger, S., Dander, A., Fischer, M., et al. (2013). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*. Available from <https://doi.org/10.1093/bib/bbs086>.
- PacBio (Pacific Bioscience). Available from <http://www.pacificbiosciences.com/>.
- Pallett, D. W., Ho, T., Cooper, I., & Wang, H. (2010). Detection of cereal yellow dwarf virus using small interfering RNAs and enhanced infection rate with cocksfoot streak virus in wild cocksfoot grass (*Dactylis glomerata*). *Journal of Virological Methods*, *168*, 223–227.
- Pantaleo, V., Saldarelli, P., Miozzi, L., Giampetruzzi, A., Gisel, A., Moxon, S., et al. (2010). Deep sequencing analysis of viral short RNAs from an infected Pinot noir grapevine. *Virology*, *408*, 49–56.
- Paszkiwicz, K., & Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in Bioinformatics*, *11*, 457–472.
- Patel, R. K., & Jain, M. (2012). NGS QC ToolKit: A toolkit for quality control of next generation sequencing data. *PLoS One*, *7*, e30619.
- Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., & Ravnika, M. (2017). Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches. *Frontiers in Microbiology*, *8*, 1998.
- Pirovano, W., Miozzi, L., Boetzer, M., & Pantaleo, V. (2015). Bioinformatics approaches for viral metagenomics in plants using short RNAs: model case of study and application to a *Cicer arietinum* population. *Frontiers in Microbiology*, *5*, 790.
- Poojari, S., Alabi, O. J., Fofanov, V. Y., & Naidu, R. A. (2013). A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family Geminiviridae implicated in grapevine red leaf disease by next-generation sequencing. *PLoS One*, *8*, e64194.
- Qi, X., Bao, F. S., & Xie, Z. (2009). Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis. *PLoS One*, *4*, e4971.
- Roossinck, M. J. (2015). Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. *Frontiers in Microbiology*, *5*(767), 10.
- Roossinck, M. J., Martin, D. P., & Roumagnac, P. (2015). Plant virus metagenomics: Advances in virus discovery. *Phytopathology*, *105*, 716–727.
- Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesnakurti, P., et al. (2017). Application of next generation sequencing for diagnostic testing of tree fruit viruses and viroids. *Plant Disease*, *101*, 1489–1499.

- Roy, A., Choudhary, N., Guillermo, L. M., Shao, J., Govindarajulu, A., Achor, D., et al. (2013). A novel virus of the Genus Cilevirus causing symptoms similar to citrus leprosis. *Phytopathology*, *103*, 488–500.
- Ruiz-Ruiz, S., Navarro, B., Gisel, A., Pena, L., Navarro, L., Moreno, P., et al. (2011). Citrus tristeza virus infection induces the accumulation of viral small RNAs (21–24 nt) mapping preferentially at the 3'-terminal region of the genomic RNA and affects the host small RNA profile. *Plant Molecular Biology*, *75*, 607–619.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5477.
- Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18.
- Seguin, J., Rajeswaran, R., Malpica-López, N., Martin, R. R., Kasschau, K., Dolja, V. V., et al. (2014). De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. *PLoS One*, *9*, e88513.
- SeqLL. (2014). *SeqLL tSMS SeqLL technical explanation*. Retrieved 12.11.20.
- Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*, 1135–1145.
- Silva, T. F., Romanel, E. A. C., Andrade, R. R. S., Farinelli, L., Osteras, M., Deluen, C., et al. (2011). Profile of small interfering RNAs from cotton plants infected with the polerovirus Cotton leafroll dwarf virus. *BMC Molecular Biology*, *12*, 40.
- Smith, O., Clapham, A., Rose, P., Liu, Y., Wang, J., & Allaby, R. G. (2014). A complete ancient RNA genome: Identification, reconstruction and evolutionary history of archaeological barley stripe mosaic virus. *Scientific Reports*, *4*, 4003.
- Soliman, T., Mourits, M. C. M., Oude Lansink, A. G. J. M., & van der Werf, W. (2012). Quantitative economic impact assessment of an invasive plant disease under uncertainty – A case study for potato spindle tuber viroid (PSTVd) invasion into the European Union. *Crop Protection (Guildford, Surrey)*, *40*, 28–35.
- Soueidan, H., Schmitt, L. A., Candresse, T., & Nikolski, M. (2014). Finding and identifying the viral needle in the metagenomic haystack: Trends and challenges. *Frontiers in Microbiology*, *5*, 739.
- Stobbe, A. H., & Roossinck, M. J. (2014). Plant virus metagenomics: What we know and why we need to know more. *Frontiers of Plant Science*, *5*(150), 10.
- Straiton, J., Free, T., Sawyer, A., & Martin, J. (2019). From Sanger sequencing to genome databases and beyond. *BioTechniques Future Science*, *66*(2), 60–63.
- Suzuki, Y. (2020). Advent of a new sequencing era: Long-read and on-site sequencing. *Journal of Human Genetics*, *65*, 1.
- Syller, J., & Grupa, A. (2016). Antagonistic within-host interactions between plant viruses: Molecular basis and impact on viral and host fitness. *Molecular Plant Pathology*, *17*, 769–782.
- Szittyá, G., Moxon, S., Pantaleo, V., Toth, G., Rusholme, P. R. L., Moulton, V., et al. (2010). Structural and functional analysis of viral siRNAs. *PLoS Pathogens*, *6*, e1000838.
- Van der Walt, A. J., van Goethem, M. W., Ramond, J. B., et al. (2017). Assembling metagenomes, one community at a time. *BMC Genomics*. Available from <https://doi.org/10.1186/s12864-017-3918-9>.
- Villamor, D. E. V., Mekuria, T. A., & Eastwell, K. C. (2016). High-throughput sequencing identifies novel viruses in nectarine: Insights to the etiology of stem-pitting disease. *Phytopathology*, *106*, 519–527.
- Visser, M., Maree, H. J., Rees, D. J., & Burger, J. T. (2014). High-throughput sequencing reveals small RNAs involved in ASGV infection. *BMC Genomics*, *15*(1), 568.
- Vives, M. C., Velazquez, K., Pina, J. A., Moreno, P., Guerri, J., & Navarro, L. (2013). Identification of a new enamovirus associated with citrus vein enation disease by deep sequencing of small RNAs. *Phytopathology*, *103*, 1077–1086.
- Wu, Q., Ding, S. W., Zhang, Y., & Zhu, S. (2015). Identification of viruses and Viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. *Annual Review of Phytopathology*, *53*, 425–444.
- Wu, Q., Wang, Y., Cao, M., Pantaleo, V., Burgyan, J., Li, W. X., et al. (2012). Homology-independent discovery of replicating pathogenic circular RNAs by deep sequencing and a new computational algorithm. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 3938–3943.
- Wylie, S. J., & Jones, M. G. K. (2011). The complete genome sequence of a Passion fruit woodiness virus isolate from Australia determined using deep sequencing, and its relationship to other potyviruses. *Archives of Virology*, *156*, 479–482.
- Xu, Y., Huang, L., Fu, S., Wu, J., & Zhou, X. (2012). Population diversity of Rice stripe virus—Derived siRNAs in three different hosts and RNAi-based antiviral immunity in *Laodelphax striatellus*. *PLoS One*, *7*, e46238.
- Xu, Y., Li, S., Na, C., Yang, L., & Lu, M. (2019). Analyses of virus/viroid communities in nectarine trees by next-generation sequencing and insight into viral synergisms implication in host disease symptoms. *Scientific Reports*, *9*, 12261.
- Yan, F., Zhang, H., Adams, M., Yang, J., Peng, J., Antoniw, J., et al. (2010). Characterization of siRNAs derived from rice stripe virus in infected rice plants by deep sequencing. *Archives of virology*, *155*, 935–940.
- Yang, X., Wang, Y., Guo, W., Xie, Y., Xie, Q., Fan, L., et al. (2011). Characterization of small interfering RNAs derived from the geminivirus/betasatellite complex using deep sequencing. *PLoS One*, *6*, e16928, 10.
- Yoshikawa, N., Yamagishi, N., Yaegashi, H., & Ito, T. (2012). Deep sequence analysis of viral small RNAs from a green crinkle-diseased apple tree. *Petria*, *22*, 292–297.
- Zahid, K., Zhao, J. H., Smith, N. A., Schumann, U., Fang, Y. Y., Dennis, E. S., et al. (2015). Nicotiana small RNA sequences support a host genome origin of cucumber mosaic virus satellite RNA. *PLoS Genetics*, *11*, e1004906.
- Zhang, C., Wu, Z., Li, Y., & Wu, J. (2015). Biogenesis, function, and applications of virus-derived small RNAs in plants. *Frontiers in Microbiology*, *6*, 1237.

- Zhang, W., Chen, J., Yang, Y., et al. (2011). A practical application of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS One*, *6*, e17915.
- Zhang, Z., Qi, S., Tang, N., Zhang, X., Chen, S., Zhu, P., et al. (2014). Discovery of replicating circular RNAs by RNA-seq and computational algorithms. *PLoS Pathogens*, *10*, e1004553.
- Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. M., et al. (2017). Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, *12*(1), 44–73.

Microbial degradation of herbicides in contaminated soils by following computational approaches

Kusum Dhakar^{1,2}, Hanan Eizenberg¹, Zeev Ronen², Raphy Zarecki^{1,2} and Shiri Freilich¹

¹*Neve Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel,* ²*Department of Environmental Hydrology & Microbiology, Zuckerberg Institute for Water Research, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel*

25.1 Herbicides: use and impact on environment

Herbicides are the nonnatural agrochemicals that are being used for the removal of unwanted plants from the crop field to enhance the agricultural yield. The majority of the herbicides kill plants through interference with the photosynthetic system. A diagrammatic overview of the attacking mechanism of the herbicide diuron is shown in Fig. 25.1. Among the pesticides, herbicides have been reported to have a very high percentage (> 45%) of sales in comparison to other categories. Herbicides can be classified on the basis of their translocation, time and method of application, and mode of actions (Vats, 2015). The other classification on the basis of chemical (considers active ingredients and mode of action) (Forouzesh, Zand, Soufizadeh, & Samadi Foroushani, 2015) properties of herbicide is also an efficient categorization to study the group effect. As we know that the growing need of food has put the agricultural system on stress to enhance crop yields, it indirectly increased the load of such agrochemicals on crop soils. Due to the consistent use and ineffective management for removal, these xenobiotics are accumulating in the environment and exerting toxic effect on the other component of the ecosystem, including humans (Bailey-serres, Parker, Ainsworth, Oldroyd, & Schroeder, 2019; Kah, 2020; Meena et al., 2020). The chemical structure of the herbicides plays an important role in their interaction and persistence in the environment. Natural removal of these xenobiotics from the environment is very slow and depends on the environmental conditions. Although the use and impact of the herbicide on environment vary, the overall measurement is required about the toxicity of the herbicide to the nontarget populations. Environmental impact quotient (EIQ) is a measure of the effect of herbicide on environment. EIQ can provide a categorization of herbicide on the basis of amount of risk associated (Kniss & Coburn, 2015).

Herbicides come from a wide variety of chemical groups that interact with environment variably. Triazine herbicides are known to have their long persistence and negative effect on the environment (Chan, Chan, & Wong, 2019). Due to extensive use of the atrazine and potential hazards to human, it has become a serious issue. Investigations for its remediation are in interest because it can contaminate the sources of drinking water (high persistence and mobility) (Mudhoo & Garg, 2011). Hazardous effect of atrazine has been recorded on a range of organisms from invertebrates to vertebrates. Toxicity of atrazine on the various systems of human body has been recorded (Singh et al., 2018). Similarly, phenyl urea herbicides (PUHs) are being used extensively to remove weeds from the crop fields. The member of PUHs (i.e., diuron, linuron, isoproturon) is being identified as a serious threat to environment. Being highly soluble in water, PUHs have been reported as a pollutant in water bodies (Hussain et al., 2015).

Glyphosate (trade name: Roundup), a postemergence herbicide, is being used globally for weed control. Aquatic environments have been found to be highly affected by the glyphosate herbicide. Bioaccumulation and food chain contamination of glyphosate are identified mainly with aquatic organisms, at some extent (Annett, Habibi, & Hontela, 2014). Glyphosate is a chelator and has the ability to bind bivalent ions (i.e., Ca⁺², Mg⁺², Mn⁺²) and exerts an inhibitory effect on enzyme involved in shikimate pathway by associating with manganese (Richmond, 2018). An intermediate, aminomethylphosphonic acid of glyphosate transformation is more stable than the herbicide itself. It is found that

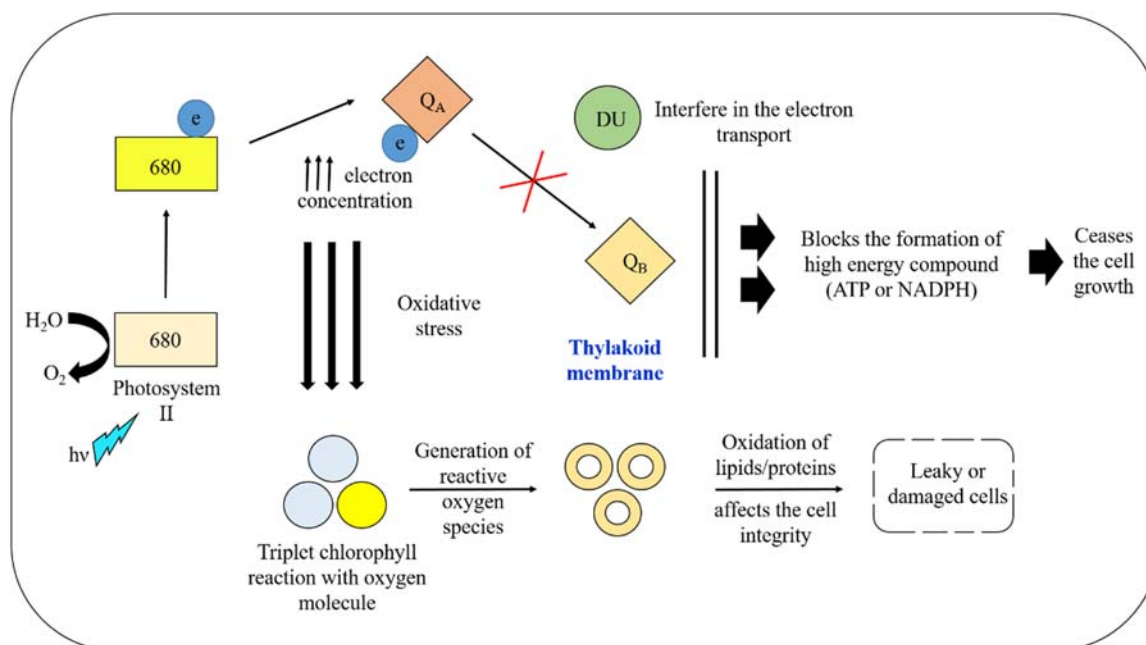


FIGURE 25.1 An overview: effect of diuron as herbicide on targeted plant (Haynes et al., 2000).

glyphosate has low impact on environment in comparison to other xenobiotics used (Duke, 2020). Herbicides (quizalofop-p-ethyl and cycloxydim) from arylphenoxypropionate group are postemergence herbicide have been recognized as a source to contaminate groundwater due to their solubility in water and can exert negative effect to the aquatic organisms (Rosculete, Bonciu, Rosculete, & Oлару, 2019).

More than five decades, 2,4-D (2,4-dichlorophenoxyacetic acid) is being used actively in agriculture and possesses a significant impact on environment (Peterson, McMaster, Riechers, Skelton, & Stahlman, 2016). The herbicide is associated with high consumption worldwide and in environment, mainly it is found as a contaminant of water due to its high solubility in water (indicating serious health hazard). Bioaccumulation of 2,4-D is reported on a faster rate and possesses toxic effects on various organisms (Islam et al., 2018). The removal of 2,4-D from the environment is an important issue due to its high toxicity on the environment (Zuanazzi, Ghisi, & Oliveira, 2020). In addition, the repeated use of the available herbicides (xenobiotics) in the crop soils increasing the herbicide resistance in the weeds and novel herbicides are needed to manage the resistant weeds. Researchers are trying to find out new targets/mode of actions that remove the unwanted weeds from the crop fields (Qu et al., 2021) which in parallel is required to develop strategy for their (residual) removal from environment.

Researchers have tried a range of approaches for the rapid removal of these pollutants from environment. Although several investigations have been done to get rid of pollutants effectively and efficiently, the majority of the researches have been carried out at small scale for a limited time duration. It is necessary to increase the in situ evaluation with detailed investigations (Sun, Sidhu, Rong, & Zheng, 2018). Several physical, chemical, and biological strategies are being followed to improve the bioremediation of herbicides (He et al., 2019; Pileggi, Pileggi, & Sadowsky, 2020; Saravanan et al., 2020; Souza et al., 2016). Among the approaches, microbial degradation is an important tool to develop an efficient strategy against the xenobiotic pollution. We know that the degradation of such synthetic compounds in environment is mainly due to the microorganisms. The cell factory of microbes can be utilized for the cleaning of environment by integrating appropriate technology into it.

25.2 Microbial degradation of herbicides

Microorganisms are considered principal degraders of the herbicide in a wide range of contaminated environments (Singh, Kuhad, Singh, Lal, & Tripathi, 1999). The consistent development in the molecular technologies helped to extend our knowledge related to the degradation potential of microorganisms (Trigo, Valencia, & Cases, 2009). In the environment, microorganisms coexist in communities and participate in the degradation process. Such degradation

processes in environment are very complicated due to the interaction of several biotic and abiotic components (Li et al., 2018).

Till date, several microorganisms from all the three domains of life have been explored for their relevance to biodegradation in environment. Among bacteria, *Pseudomonas* is a genus widely known for its ability to degrade a range of synthetic substances. *Pseudomonas fluorescens* was extensively investigated at molecular level for degradation pathway of sulfonylurea (Zanardini et al., 2002). In last more than 20 years, *Pseudomonas* sp. strain ADP (Adenosine di phosphate) has been investigated for atrazine degradation potential. Strain ADP harbors the atrazine degradation pathway and is able to utilize atrazine as sole nitrogen source (Boundy-Mills, De Souza, Mandelbaum, Wackett, & Sadowsky, 1997; Esquirol et al., 2018). In addition to *Pseudomonas*, *Arthrobacter* is one of the efficient degraders of atrazine. Several strains of *Arthrobacter* have been investigated for its potential to degrade atrazine under different environmental conditions (Aislabie, Bej, Ryburn, Lloyd, & Wilkins, 2005; Wang & Xie, 2012). There is a diversity in atrazine degradation by *Arthrobacter*. TC1 strain of *Arthrobacter* is reported to transform atrazine in cyanuric acid (Strong, Rosendahl, Johnson, Sadowsky, & Wackett, 2002) but some other strains of *Arthrobacter* have also been identified for the utilization of cyanuric acid as nitrogen source (Hatakeyama et al., 2015).

In addition to *Arthrobacter*, *Kaistobacter* (Lin et al., 2018), *Acinetobacter* (Yang, Jiang, Zhu, Zhao, & Zhang, 2017), *Bacillus* (Khatoun & Rai, 2020), and *Citricoccus* (Yang, Wei, Zhu, & Geng, 2018) are some of the bacterial genera identified as efficient degraders of atrazine. Recently, *Fusarium*, a fungus, is reported for its ability to degrade atrazine at some extent (Esparza-Naranjo et al., 2020). Though some of the bacteria have been considered a potential candidate against atrazine pollution, still the need of identifying new degraders is required. *Bacillus licheniformis* and *megaterium* spp. have been reported for their ability to degrade atrazine. Both the strains of *Bacillus* show faster degradation in consortia in comparison to degradation in isolated culture (Zhu, Fu, Jin, Meng, & Yang, 2019).

PUHs are one of the most important categories in herbicide which is widely used to remove weeds or unwanted plants. Diuron, linuron, isoproturon are some of the important herbicides among the phenyl urea category. Diuron degradation in soil is mainly the outcome of microbial activity along with a small fraction of photochemical decomposition (Kovács et al., 2016; Tasca & Fletcher, 2018). The rate of degradation is affected by physicochemical factors such as soil pH, soil texture, and available organic matter (Guimarães et al., 2018). Biological factors affecting the process include the crop plants that can stimulate the process (Piutti, Marchand, Lagacherie, Martin-Laurent, & Soulas, 2002). Additional factors include the quality and quantity of the soil organic matter directly affecting the sorption of the diuron and its accessibility to the microbes. Finally, the high amount of aromatic content in the organic matter enhances the binding of diuron with the soil (Albers, Banta, Hansen, & Jacobsen, 2008).

Decomposition process typically leads to the formation of 3,4-DCA (DCA), a highly persistent pollutant that is more toxic than diuron. The information on the complete mineralization pathway of the 3,4-DCA is still scarce. Although several bacteria (including *Streptomyces*) and fungi have been reported to degrade the highly stable metabolite, still the complete pathway needs more investigations (Arora, 2015; Briceño, Fuentes, Saez, Diez, & Benimeli, 2018; Giacomazzi & Cochet, 2004). The capability of *Pseudomonas* to degrade DCA was suggested through aromatic ring cleavage and its utilization as a carbon source (El-Deeb, Soltan, Ali, & Ali, 2000). On the same line, the strong evidences for the DCA degradation in *Pseudomonas* have been provided by Kim et al. (2007). The presence of 12 genes linked to catechol pathway (including catechol 2,3-dioxygenase) was revealed for DCA degradation (similar enzyme in *Pseudomonas acidovorans* was reported earlier also Hinteregger, Loidl, & Streichsbier, 1992). You and Bartha (1982) observed muconate and butenolide which may likely have a fate toward the oxo adipate via maleylacetate. In a study the transformation of DCA in *Acinetobacter* showed the formation of aniline and 4-chlorocatechol followed by the ortho cleavage mechanism (Hongsawat & Vangnai, 2011). On the other hand, 1, 2 catechol dioxygenase gene was recognized as a part of the phenyl urea degradation by *Sphingobium* sp. (Sun et al., 2009).

Fungi have been considered an efficient degrader in soil. The fungi are also recognized for their efficient enzyme system for the xenobiotic degradation. The various oxido-reductases and peroxidases have been found to be involved in a range of mineralization of aromatic hydrocarbons (Spina et al., 2018). The involvement of oxidases along with the antioxidants has been demonstrated in the *Ganoderma lucidum* (Coelho-Moreira et al., 2018). Diuron degradation by *Mortierella* sp. suggests the formation of the nonaromatic diol followed by N-dealkylation (Badawi et al., 2009). Similarly, N-dealkylation-mediated degradation process has also been reported in *Neurospora intermedia*, an endophyte, isolated from the sugarcane root (Wang, Li, Feng, Du, & Zeng, 2017). Apart from these proteins/enzymes, N-acetyltransferase is found in fungi that is responsible for the acetylation of 3,4-DCA and results in the less toxic product (Martins, Dairou, Rodrigues-Lima, Dupret, & Silar, 2010). The consortia of *Aspergillus brasiliensis* and *Cunninghamella elegans* showed the significant transformation of diuron to reduce the toxicity of the metabolites. The major intermediates observed through liquid chromatography were DCPMU (1-(3,4-dichlorophenyl)-3-methylurea),

DCPU ((3,4-dichlorophenyl)urea), DCA along with the 3,4-dichloro acetanilide (Perissini-Lopes et al., 2016). The production of acetanilides have also been observed in the degradation studies of DCA by bacteria (Egea et al., 2017; Yao et al., 2011).

For linuron degradation, *Variovorax* is highly investigated at genomic and proteomic level. *Variovorax* sp. strain SRS16 can utilize linuron as sole carbon and nitrogen source, the strain is also explored for its degradation pathway and being considered a promising candidate for bioremediation of linuron (Bers et al., 2011; Öztürk et al., 2020; Sørensen, Simonsen, & Aamand, 2009). A synergistic catabolism of linuron has been reported by *Diaphorobacter* and *Achromobacter*, where the first one carries out the initial hydrolysis of linuron and second one mineralizes the produced aniline derivatives (Zhang, Hang, et al., 2018). Isoproturon was found to be rapidly mineralized by *Sphingomonas* sp., isolated from a contaminated agricultural soil (Hussain, Sørensen, Devers-Lamrani, El-Sebai, & Martin-Laurent, 2009). *Pseudomonas aeruginosa* strain JS-11 was used as an efficient bioinoculant against isoproturon pollution. The strain has a positive contribution not only with bioremediation but with plant growth and disease management aspects too (Dwivedi, Singh, Al-Khedhairy, & Musarrat, 2011). By following the green chemistry, fungal enzymes mainly laccase (oxidoreductase) have been identified as a potential tool to remove isoproturon from the contaminated sites (Zeng, Qin, & Xia, 2017).

Being in consistent use and stable compound, nonnatural phosphonates have become a serious environmental issue. Glyphosate is a representative member of phosphonate and recognized as a contaminant of soil and water (Sviridov et al., 2015). In bioremediation of glyphosate, *Bacillus subtilis* strain has been reported to have a significant ability to remove the contaminant (Yu et al., 2015). *Comamonas odontotermitis* P2 strain can be found to have the ability to utilize glyphosate as carbon and phosphorous source and can be a potential candidate for environmental removal of glyphosate (Firdous, Iqbal, & Anwar, 2020). Not only bacteria but also fungi are also known for efficient degradation of glyphosate. A fungal strain of *Trichoderma viride* can utilize glyphosate as a sole phosphorus source and can be a promising candidate in the bioremediation of glyphosate (Arfarita et al., 2013). In addition to *Trichoderma*, *Aspergillus oryzae* also reported to be an efficient degrader of the herbicide (Carranza et al., 2019).

The rapid development in the molecular and computational methods allows researchers to find out new and effective solutions for bioremediation. A strain of *Paenibacillus polymyxa* has been reported to possess degradation activity for five different xenobiotics (Zhang et al., 2019). Such efficient strains need to be explored extensively related to degradation process for further utilization at commercial scale in bioremediation. New microbial degraders are being identified and investigations to reveal their metabolic abilities and their interactions in communities are going on. A considerable impact of the mixed culture of *Pseudomonas* and *Achromobacter* on herbicide degradation was recently investigated (Yang et al., 2020). Some examples of the microbial degraders of various herbicides have been included in Table 25.1.

25.3 Strategies to improve biodegradation of herbicides

Although herbicides are considered a potent tool in the service of the agriculture economy, due to the excess and consistent use, the fate of the herbicide has been found to exert negative impacts on environment. The developing methods through science and technology are widely supporting to remove the pollutants from the soil and water. Several research groups are working to develop methods that facilitate the bioremediation. Most of these methods are combined with the degrading microorganisms/either pure culture or in consortium (Geed, Prasad, Kureel, Singh, & Rai, 2018; Santos et al., 2019). Microbial degradation in vitro and in vivo is a result of interactions of several parameters associated to the growth conditions. The degradation process can be enhanced by evaluating the effects of individual parameters and the stability of the intermediates during the process (Wang, Lai, Latino, Fenner, & Helbling, 2018).

Several strategies have been followed to enhance the biodegradation of herbicides in the various environments (Fig. 25.2). In bioremediation, bioaugmentation and biostimulation are considered an effective approach. In bioaugmentation, addition of living cells (degraders) is included for the rapid removal of pollutants from the environment (Adams, Fufeyin, Okoro, & Ehinomen, 2015). Bioaugmentation is a strategy to enhance the degradation capacity of the polluted environment by introducing efficient exogenous degraders. There is a plethora of information on the microbial degradation of xenobiotic substances. *Arthrobacter*, *Bacillus*, *Burkholderia*, and *Pseudomonas* are some of the bacterial genera renowned to have ability to degrade a range of nonnatural substances (Singh & Singh, 2016). The extensively characterized microorganisms for degradation are being utilized at a wide scale under bioaugmentation process (Cycoń, Mrozik, & Piotrowska-Seget, 2017). For example, The impact of *Pseudomonas* sp. ADP strain on the triazine degradation in soil as a bioaugmented bacteria was found to be weak (Morán, Müller, Manzano, & González, 2006) whereas bioaugmentation of *Pseudomonas* MHP41 was reported to have a significant impact on the degradation and soil microbial community in the contaminated soil (Morgante et al., 2010).

TABLE 25.1 Examples of herbicide degradation by microorganisms.

S. no.	Herbicide	Degrader (Scientific name)	Classified as	References
1	2,4-Dichlorophenoxyacetic acid	<i>Acinetobacter</i> sp., <i>Stenothrophomonas maltophilia</i> , <i>Flavobacterium</i> , <i>Serratia marcescens</i> , and <i>Penicillium</i> sp.	Bacteria Fungi	Silva et al. (2007)
2	Diuron	<i>Micrococcus</i> sp.	Bacteria	Sharma and Suri (2011)
3	Oxyfluorfen	<i>Pseudomonas</i> sp., <i>Arthrobacter</i> spp., <i>Mycobacterium</i> sp., <i>Micrococcus</i> sp., <i>Streptomyces</i> sp., and <i>Aspergillus</i> sp.	Bacteria Fungi	Mohamed, El Hussein, El Siddig, and Osman (2011)
4	Chloroacetamide	<i>Paracoccus</i> sp.	Bacteria	Zhang et al. (2011)
5	Acetochlor (chloroacetamide)	<i>Rhodococcus</i> sp., <i>Delftia</i> sp. and <i>Sphingomonas</i> sp. (consortium)	Bacteria	Hou et al. (2014)
6	Atrazine	<i>Anthracytophyllum discolor</i>	Fungi	Elgueta, Santos, Lima, and Diez (2016)
7	Triazine	<i>Leucobacter</i> sp.	Bacteria	Liu et al. (2017)
8	2,4-Dichlorophenoxyacetic acid	<i>Cupriavidus gilardii</i>	Bacteria	Wu et al. (2017)
9	2,4-D	<i>Umbelopsis isabellina</i>	Fungi	Bernat et al. (2018)
10	Nicosulfuron	<i>Plectosphaerella cucumerina</i>	Fungi	Carles et al. (2018)
11	Diuron	<i>Ganoderma leucidum</i>	Fungi	Coelho-Moreira et al. (2018)
12	Alachlor	<i>Trichoderma koningii</i>	Fungi	Nykiel-Szymańska, Bernat, and Słaba (2018)
13	Chlorimuron-ethyl	<i>Enterobacter ludwigii</i>	Bacteria	Pan, Wang, Shi, Fang, and Yu (2018)
14	Alachlor (chloroacetamide)	<i>Xanthomonas axonopodis</i> , <i>Aspergillus niger</i> , <i>A. flavus</i> , and <i>Penicillium chrysogenum</i>	Bacteria, fungi	Ahmad (2020)
15	Diuron, sulfentrazone, 2,4-D, and oxyfluorfen	<i>Bradyrhizobium</i> sp.	Bacteria	Madureira Barroso et al. (2020)
16	Atrazine	<i>Bjerkandera adusta</i>	Fungi	Dhiman et al. (2020)
17	Butachlor	<i>Bacillus altitudinis</i>	Bacteria	Kaur and Goyal (2020)
18	Atrazine	<i>Pleurotus ostreatus</i>	Fungi	Lopes et al. (2020)
19	Atrazine	<i>Pseudomonas</i> sp., <i>Arthrobacter</i> sp., <i>Variovorax</i> sp., <i>Chelatobacter</i> sp.	Bacteria	Billet et al. (2021)
20	Butralin (dinitroaniline)	<i>Sphingopyxis</i> sp.	Bacteria	Ghatge et al. (2021)

Against triazines, *Arthrobacter* has been considered a promising candidate for the bioaugmentation since it is known to have a range of catabolic pathways for xenobiotic degradation (Sagarkar et al., 2016; Xu et al., 2019). *Novosphingobium* was identified as a promising bioaugmented bacterial strain in the contaminated soils with herbicide 2,4-dichlorophenoxyacetic acid (2,4-D) (Dai, Li, Zhao, & Xie, 2015). Two bioaugmented bacterial species *Pseudochrobactrum* and *Masilia* were reported to have a significant impact on the degradation process of chlorothalonil in soil (Xu et al., 2018). A nitrile degrading bacterium *Rhodococcus rhodochrous* is extensively explored for its ability to degrade nitrile and proposed as a promising candidate for bioaugmentation in the contaminated sites with nitriles

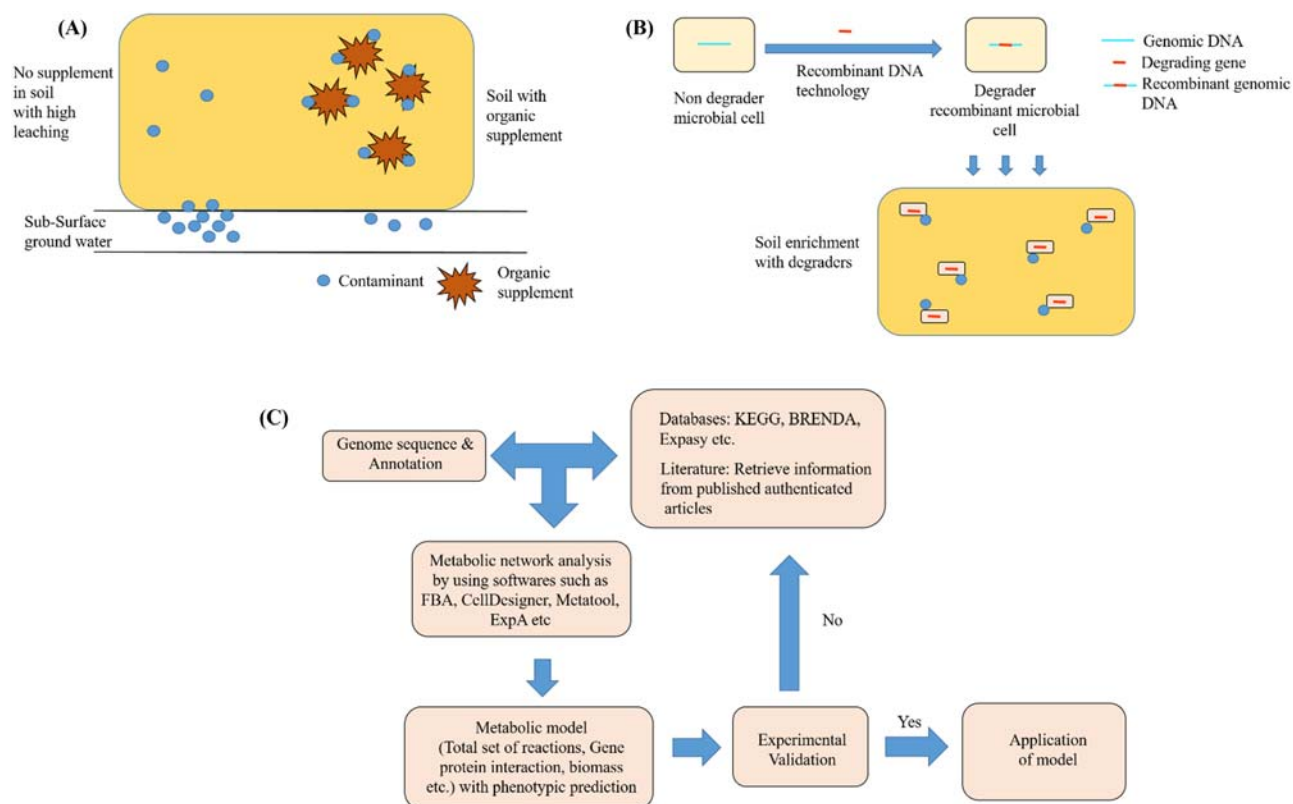


FIGURE 25.2 Different strategies to improve biodegradation in soil. (A) Addition of amendments/biostimulants to alter the absorption and bioavailability of herbicides in contaminated environment. (B) Construction and introduction of recombinants to enhance degradation process. (C). A generic scheme for computational-based strategies for the optimization of process in bioremediation.

(An et al., 2020). Although bioaugmentation strategy is considered an efficient approach, the introduction of exogenous microbial species might bring some uncalculated effects associated with the indigenous microbial community (Pacwa-Płociniczak, Płaza, & Piotrowska-Seget, 2016). To enhance the efficiency of the bioaugmented process and maintain the interactions of the bioaugmented species, the optimization processes play a significant role in the process (Valdez-Vazquez, Castillo-Rubio, Pérez-Rangel, Sepúlveda-Gálvez, & Vargas, 2019).

Besides bioaugmentation, biostimulation is recognized as another efficient approach deals with the stimulation of indigenous microbial degraders to enhance the degradation rate in contaminated environment (Kanissery & Sims, 2011). In the bioremediation studies, generally eco-friendly biostimulants have been a priority of the researchers to improve the soil health. Agri and food wastes can be utilized as an efficient biostimulant after the optimization process (Xu & Geelen, 2018). The addition of organic waste changes the physical and nutritional status of the soil and activates indigenous microorganisms for rapid degradation (Briceño, Palma, & Durán, 2007). A study suggests that soil amended with compost and corn-related organic waste influenced the atrazine degradation rate in soil. A significant enhancement in bacterial load and dehydrogenase enzyme activity was recorded in relation to high degradation rate (Moorman, Cowan, Arthur, & Coats, 2001). Similar results, high microbial activity with dehydrogenase was detected in the bioremediation process under the influence of Olive cakes as organic amendment (Delgado-Moreno & Peña, 2007). A study suggested the role of soil type in the final impact of the amendments on degradation process. The different degradation rate of herbicide was recorded in the sandy loam and silty clay soils (Forouzangohar, Haghnia, & Koocheki, 2005). The effect of organic and inorganic amendment was investigated by Kadian, Gupta, Satya, Mehta, and Malik (2008) on atrazine degradation. Among the tested ones the highest atrazine degradation rate was recorded with the biogas slurry amended conditions.

Manure with sodium citrate (inorganic amendment) showed strong enhancement after a lag. Farmyard manure in comparison to rice straw, sawdust, and compost was reported to have stronger effect on the degradation rate of atrazine (Mukherjee, 2009). It is reported that some organic amendments show a negative effect on the degradation process. It is already known that in general the microbial load and the functional activity (e.g., dehydrogenase activity) enhance due

to the addition of organic amendments indicating high rate of degradation. The reverse effect of oak and pinewood-based amendments was recorded on linuron degradation in soil. The wood-based amendments possibly adsorbed the linuron strongly and reduced its bioavailability which led to high persistence of linuron in comparison to nonamended soils (Grenni, Caracciolo, Rodríguez-Cruz, & Sánchez-Martín, 2009). It has been noticed that same soil amendment impacts differently on the degradation of different herbicide. Oak- and pine-based amendments showed a significant contribution in the degradation of terbuthylazine in contaminated soil. These amendments alter the rate of microbial degradation in the soil by manipulating the adsorption and bioavailability of the herbicide in contaminated soils along with limiting its mobility (Grenni et al., 2012). Enhancement of herbicide retention capacity of soils can be achieved through the organic amendments, a potential strategy to make the pollutant available for microbes to degrade. This also reduces the spread of the herbicide in the various components of environment through processes like leaching (Gámiz, Celis, Hermosín, & Cornejo, 2010). The microbial response to the herbicide and the soil amendments varies with time and the amount of herbicide in soil. The dissipation of mesotrione was assessed in the amended and unamended soil with a range of doses which recruited a significant effect on the microbial load and functional activity (Pose-Juan, Sánchez-Martín, Herrero-Hernández, & Rodríguez-Cruz, 2015).

Organic carbon processed through a different physical process impact differently as amendment in soils. The degradation process of isoproturon was variably affected by pyrochar and hydrochar, the later one was comparatively better for the herbicide accessibility to microbial degradation (Eibisch, Schroll, & Fuß, 2015). In general, biochar alters the soil properties and affects the adsorption of the herbicides which leads to the change in the microbial degradation rate of herbicide in soil (Zhelezova, Cederlund, & Stenström, 2017).

Green compost was identified as an effective amendment in soil to increase the persistence of prosulfocarb and trisulfuron and significantly affected leaching process (Marín-Benito, Barba, Ordax, Sánchez-Martín, & Rodríguez-Cruz, 2018). In diuron removal, plant husks (natural adsorbents) are being used due to its good adsorption ability and fast removal of contaminant from water (Bezerra et al., 2020). The addition of organic matrix to the soil is more eco-friendly approach that can facilitate the degradation of the herbicides. The main problem with the herbicide diuron is that its accessibility to the microbes is very less. The cyclodextrin (hydroxypropyl- β -cyclodextrin) disrupts the strong bonding between diuron and organic matter that results in high solubility of diuron. The consortium of degrading bacteria further mineralizes the herbicides to CO₂. The strategy showed positive results in achieving a high degree of degradation in the contaminated soil system (Rubio-Bellido, Morillo, & Villaverde, 2018).

Biochar (product of partial pyrolysis of wood/organic matter) enhances the soil fertility by altering the soil physico-chemical properties and the associated microbiota (Liu, Lonappan, Brar, & Yang, 2018). The amendment of soil by mixing the biochar contributed positively to biodegradation. It has the ability to adsorb the xenobiotic compounds and accelerate their dissipation process (Zhelezova et al., 2017). Further, the rate of degradation of the adsorbed molecules depends on the dynamics of interactions between soil and biochar (Rubio-Bellido et al., 2018). Sometimes, the physical factor (e.g., temperature) dominates the effect of organic amendments. The increase in temperature accelerates the microbial activity that led to a rapid degradation in the soil (Marín-Benito, Carpio, Sánchez-Martín, & Rodríguez-Cruz, 2019).

In addition, nanobiotechnology is an advanced branch of biotechnology and has been proved its worth in remedy of various problems in environment and health science. The nanomaterials have showed favorable effect on the phytoremediation and bioremediation process (Vázquez-Núñez, Molina-Guerrero, Peña-Castro, Fernández-Luqueño, & de la Rosa-Álvarez, 2020). The removal of the pollutants from the soil through nanomaterials is considered significant. Although it has shown good results on the lines of bioremediation, still it is important to investigate its effect on the soil and the other components (biological and nonbiological) of ecosystem (Gong et al., 2018). Though all the approaches are very effective, the optimization process demands high consumption of resources, time, and efforts. To reduce the complexity and to improve the efficiency, integration of computational biology is being identified as an efficient approach.

The involvement of upcoming *in silico* methods (sequencing, metabolic models, etc.) is also recognized as a potent tool to study the microbial community response against the herbicide exposure. The metabolic modeling is an applied approach to predict the phenotype of a model on the basis of computational algorithms (Faust, 2019; Henry et al., 2010). The approach is alluring to the various scientists and engineers because you need not to play with the genomes at molecular experiments but with the sequences through computational process. This provides an immense support to produce an efficient degrader that reduces the time and costs of wet laboratory efforts (Ali, Khan, Li, Zheng, & Yao, 2019; Xu et al., 2019). The only thing that is challenging in the predictive biology is to design the consortia and the conditions to get the favorable on-ground results. The involvement of computational-based methods can increase the efficiency of bioremediation approaches (bioaugmentation or biostimulation).

25.4 Integration of computational biology to improve biodegradation of herbicides

So far we have discussed traditional and advanced methods to improve the biodegradation of herbicides in various contaminated environments. Computational-based methods are one of the efficient approaches to improve biodegradation with cost-effective and time-saving properties (Finley, Broadbelt, & Hatzimanikatis, 2009). Among the computational methods, genome-scale metabolic modeling is proven as a promising tool that helps to predict the phenotype of the microorganisms under a specific environmental condition (Terzer, Maynard, Covert, & Stelling, 2009).

Metabolic modeling helps to explore the metabolic potential of the microbial degraders and to develop improved bioremediation strategies. Either bioaugmentation or biostimulation, the efficiency of both the approaches can be enhanced by incorporating the modeling methods. The algorithms of modeling methods calculate the impact of a specific environmental condition and predict the rate of degradation *in silico*. The screening of multiple conditions extracts comparatively suitable conditions for the improved degradation process. In brief, the affected group of the microorganisms in the polluted environment is identified through high-throughput methods, that is, sequencing technologies (Jo, Oh, & Park, 2020). The identified microbial groups are then explored at genomic and metabolic level. The entire information of a particular microbe or microbial group can be retrieved through literature and online resources. The available genome and metabolic information are then integrated into modeling algorithms to study the potential of the microbial group. The predictive biology identifies the metabolic capacities of the target microbial group which can be further validated through laboratory/field experiments (Fig. 25.3).

Genome-scale metabolic model (GSMM) construction follows the integration of genomic and metabolic information with mathematical model to predict the behavior of organism. In short, GSMM refers to a mathematical framework which is constructed by incorporating all the gene proteins reaction associations based on the genome annotations and the metabolic information available (Gu, Kim, Kim, Kim, & Lee, 2019; Thiele & Palsson, 2010). To construct a high-quality genome-scale model, online available resources and related packages are useful at different stages to construct functional GSMM. Model SEED is an online resource for a rapid generation and optimization of GSMM (Henry et al., 2010). The consistent development in technology led us to several online resources other than SEED for rapid and

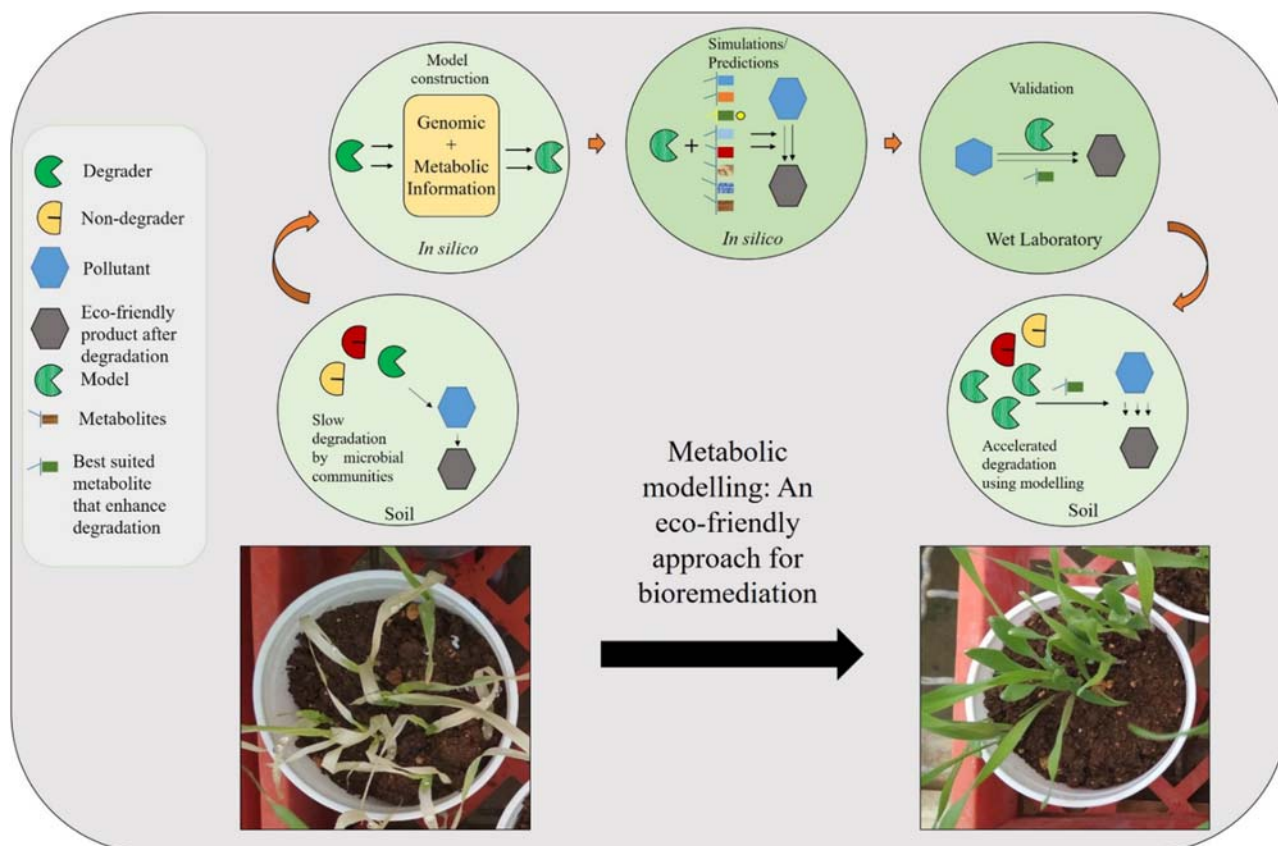


FIGURE 25.3 An overview of the different steps in metabolic modeling approach for bioremediation.

efficient construction of genome-scale models, such as Raven Toolbox, Pathway Tools (such as CarveMe) (Faria, Rocha, Rocha, & Henry, 2018). Further, the Model SEED-generated reconstruction requires biomass reaction generation and curation for a valid working metabolic model (Devoid et al., 2013). Constraint-based reconstruction and analysis (COBRA) is a modeling approach that helps in the prediction of a behavior of an organism. The conversion of fluxes in a working model can be computed by COBRA toolbox (Becker et al., 2007). BiGG knowledge base helps to integrate all the information (genomic and metabolic) of an organism systematically (Schellenberger, Park, Conrad, & Palsson, 2010).

The biomass reaction (it includes all the biomass constituents) generation plays an important role in validating the genome-scale models. To enhance the accuracy of the model, biomass objective function considered an important factor that includes the principal biomass components (nucleic acids, lipids, species-specific components, etc.) (Lachance et al., 2019). In addition, the quality and accuracy of the model also depends on the genome coverage of the model. Genome coverage is the ratio of the genome protein reaction association to the total genes present in the genome. To date, several software platforms have been developed to achieve more accurate and functional metabolic models to keep the pace with the genome sequences obtained from high-throughput sequencing technologies (Mendoza, Olivier, Molenaar, & Teusink, 2019). In this way the predictive biology contributes to identify desirable conditions through algorithms that can be validated further through wet laboratory experiments.

Several microbial GSMMs have been constructed and utilized in a range of different applications, including degradation of xenobiotics. *Pseudomonas* genus is widely known for its wide biotechnological applications. iJN746, a genome-scale model of *Pseudomonas putida* KT2440 was constructed by following COBRA approach. The metabolic capacities of the strain KT2440 was explored efficiently toward its application in biotechnology (Nogales, Palsson, & Thiele, 2008). The metabolic model PpuMBEL1071 of *P. putida* KT2440 was constructed and explored for its ability to degrade a range of aromatic compounds. The modeling also helped to extend the information on anaerobic survival of the strain through predictive biology (Sohn, Kim, Park, & Lee, 2010). In addition to the biodegradation of pollutants, phthalates are recognized as harmful xenobiotics from the plastic industries. A *Rhodococcus* strain HS-D2 reported to utilize n-butyl benzyl phthalate as a sole carbon source. The potential of bacterial strain to degrade BBP (blood-brain barrier penetration) was extensively studied in silico and in vitro. The constructed genome-scale model iYZ1601 can be utilized further to predict the degradation of capacities of the *Rhodococcus* in a wide range of environmental conditions (Zhang, Chen, et al., 2018).

Alphaproteobacterium, *Sphingopyxis granuli* strain TFA extensively explored for the complete mineralization of aromatic hydrocarbon tetralin. Genome strain metabolic model for TFA strain was constructed to explore its metabolic capacities through modeling approach. The model-based predictions were successful to reveal the consumption of new substrates as a sole carbon and energy source by TFA strain. This study helped in the unraveling of metabolic potential of the oligotrophic strain TFA (García-Romero, Nogales, Díaz, Santero, & Floriano, 2020).

The advancement in sequencing technologies and annotation tools has supported genome-scale metabolic modeling significantly (Mahadevan & Henson, 2012). Investigations unraveled the microbial potential of several organisms, including three domains of life (archaea, bacteria, and fungi). Metabolic modeling-based predictions are not only efficient in the functional predictions of single GSMMs but they work equally good to the microbial communities in the environment (Biggs, Medlock, Kolling, & Papin, 2015). It is a well-known truth that the microbial processes in environment are very complicated due to the interactions of several factors. To understand these microbial processes at molecular level, it is very important to explore their metabolic exchanges. Members of microbial communities interact (positive or negative interactions) each other to maintain life processes. Prediction of microbial community behavior is a challenging task due to a complex network of dependent and independent members and requires enough data to study community interactions. More accurate GSMMs and high-quality genomes can enhance the efficiency of community modeling (Henry et al., 2016).

Investigations on microbial interactions are useful to decode the mechanism of a process at molecular level. The enhanced cellulose mineralization by the two species of *Clostridium* in coculturing was investigated through metabolic models and extended the information on mechanism of community dynamics (Salimi, Zhuang, & Mahadevan, 2010). Another study suggested the use of constraint-based modeling of microbial community to investigate the process of oil degradation. The computationally derived metabolic fluxes along with online available resources and metagenomic data were utilized systematically to enhance the information on the oil degrading community (Röling & Van Bodegom, 2014). RedCom is an approach implemented to analyze community process based on community modeling. It included the construction of balanced microbial communities by integrating reduced stoichiometric models. It was based on the final conversions of individual species models. Here, community of nine members was investigated for the degradation in anaerobic digestion under biogas formation (Koch et al., 2019). The community process for methane metabolism was

TABLE 25.2 Some examples of genome-scale metabolic models with biotechnological applications.

S. No.	Microorganism	Classified as	Application	References
1	<i>Penicillium chrysogenum</i>	Fungi	Penicillin production	Agren et al. (2013)
2	<i>Methanococcus jannaschii</i>	Archaea	Metabolic pathways	Tsoka, Simon, and Ouzounis (2004)
3	<i>Methanosarcina barkeri</i>	Bacteria	Methane metabolism	Feist, Scholten, Palsson, Brockman, and Ideker (2006)
4	<i>Pseudomonas putida</i>	Bacteria	Biotechnological applications	Puchalka et al. (2008)
5	<i>Bacillus subtilis</i>	Bacteria	Biotechnological applications	Henry, Zinner, Cohoon, and Stevens (2009)
6	<i>Mycoplasma genitalium</i>	Bacteria	Biotechnological applications	Suthers et al. (2009)
7	<i>Synechocystis</i> sp.	Cyanobacteria	Photobiological cell factories	Montagud, Navarro, Fernández de Córdoba, Urchueguía, and Patil (2010)
8	<i>Chromohalobacter salexigens</i>	Bacteria	Physiology	Ates, Oner, and Arga (2011)
9	<i>Clostridium ljungdahlii</i>	Bacteria	Biotechnological applications	Nagarajan et al. (2013)
10	<i>Lactococcus lactis</i>	Bacteria	Dairy industry	Flahaut et al. (2013)
11	<i>Methanococcus maripaludis</i>	Archaea	Metabolic pathways	Goyal, Widiastuti, Karimi, and Zhou (2014)
12	<i>Pseudoalteromonas haloplanktis</i>	Bacteria	Biotechnological applications	Fondi et al. (2015)
13	<i>Methylobacterium buryatense</i>	Bacteria	Methane metabolism	Torre et al. (2015)
14	<i>Mortierella alpina</i>	Fungi	Arachidonic acid production	Ye et al. (2015)
15	<i>Caenorhabditis elegans</i>	Nematoda	Animal physiology	Safak Yilmaz and Walhout (2016)
16	<i>Cordyceps militaris</i>	Fungi	Cordycepin production	Vongsangnak et al. (2017)
17	<i>Yarrowia lipolytica</i>	Fungi	Lipid production	Wei, Jian, Chen, Zhang, and Hua (2017)
18	<i>Methylococcus capsulatus</i>	Bacteria	Methane metabolism	Lieven et al. (2018)
19	<i>Chromohalobacter salexigens</i>	Bacteria	Ectoine production	Piubeli et al. (2018)
20	<i>Pseudomonas aeruginosa</i>	Bacteria	Pharmacological Research	Zhu et al. (2018)
21	<i>Geobacillus icigianus</i>	Bacteria	Biotechnological applications	Kulyashov, Peltek, and Akberdin (2020)
22	<i>Lachancea kluyveri</i>	Fungi	Biotechnological applications	Nanda, Patra, Das, and Ghosh (2020)

investigated through metabolic modeling approach. High-quality models of methanotrophs were constructed and their interactions were studied as the modeled community. *Methylomonas* and *Methylobacter*, community members showed a competition over methane utilization. The study provided significant results to understand the methane utilizing mechanism in a community (Islam, Le, Daggumati, & Saha, 2020). GSMM has also been found to be an efficient tool in various applications. Some examples are shown in Table 25.2.

25.5 Bioremediation of atrazine by following metabolic modeling method

Atrazine is a widely known herbicide in consistent use for the removal of weeds in crop fields (Mueller et al., 2017). Due to the extensive use, mobility and shelf life of atrazine emerged as a serious environmental pollution. Several research found that atrazine is impacting negatively on the nontargeted population, including human (Jablonowski, Schäffer, & Burauel, 2011; Mueller et al., 2017; Singh et al., 2018). The integration of modeling approach to the study of microbial degradation of atrazine is being seen as a promising approach to improve the rate of the process in environment. *Arthrobacter* is known as an efficient degrader of atrazine and it can utilize atrazine as sole nitrogen source. *Arthrobacter* is considered a suitable candidate for model-based design of bioremediation strategies against atrazine pollution due to its available information associated with genome and metabolic pathways.

A genome-scale model of *Paenarthrobacter aurescens* TC1 was constructed to explore its potential toward atrazine degradation (Ofaim et al., 2020). The model iRZ1179 was constructed in a semiautomated manner by using Model SEED. The draft model was further improved for atrazine degradation pathway by manual curation utilizing online resources with literature. The performance of final experimentally validated model was simulated through dynamic flux balance analysis for atrazine degradation under different carbon and nitrogen sources (separately). The impact of amino acids, glucose, and phosphate was investigated in the study. The observations supported the predictions and led to an optimized condition for atrazine degradation by *P. aurescens* TC1. The study revealed that the modeling methods can be considered an efficient tool to improve the atrazine degradation through biostimulation strategies.

Biostimulation is one of the efficient approaches for the enhanced rate of herbicide degradation in soil. The integration of modeling methods with the biostimulation strategies can improve the impact of the process (Mehdizadeh et al., 2019). A case study by Xu et al. (2019) related to the improvement of atrazine degradation in crop soils by studying the community dynamics through modeling approach is included here. In the study the atrazine exposed soils were analyzed for community shifts through 16S rRNA gene amplicon sequencing. Differentially affected bacterial groups *Arthrobacter* (as degrader) and other nondegraders species were selected for the community modeling. GSMMs were constructed for the differential abundant bacterial species (degrader and nondegraders). Initial draft model was constructed by using Model SEED (Faria et al., 2018). RAST (Rapid Annotations using Subsystems Technology) algorithms were used for the genome annotation (Overbeek et al., 2014). KBase (www.kbase.us) was used for the generation of draft metabolic model. All the draft models were further manually curated on the basis of literature and online available resources such as KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016), UniProt (D506-D515, 2019), JGI (Joint Genome Institute) (Grigoriev et al., 2012) for the biochemical and physiological characteristics.

The manual curation generally involved a few important steps such as addition of new reactions on the basis of literature and genome annotations, conversion of all the reaction IDs to the standard KBase IDs, including the correction of stoichiometry and reversibility of all the reactions and elimination of futile loops. The final working model was consistent with reported experimental evidences for growth requirements of the five modeled species. Finally, all the models were combined as a single dynamic model. Dynamic flux balance analysis (Henson & Hanly, 2014) was used to simulate the growth of modeled community under the given media condition with time. Community modeling function revealed that exchange fluxes between community members enhanced the efficiency of atrazine degradation comparison to the degradation activity of the main degrader per se. A range of consortia composed of different combinations of community members were designed and further used for simulating the corresponding performance of atrazine degradation and growth under the given conditions. The simulations were validated in *in vitro*. In addition, the impact of glucose on the atrazine degradation by the modeled community was also simulated and validated in the pot experiment.

In general, optimization process in bioremediation is a hit and trial process. A range of conditions has to be screened to identify the promising ones. But the predictions related to the degradation behavior of the microbial communities under a set of given environmental conditions reduce the time and cost and improve the efficiency of optimization process. The approach is comparatively reliable because it takes all the genomic and metabolic information (molecular level interaction) in account to reach a final result.

25.6 Conclusion

The consistent advancement in the science and technology has opened new ways to improve bioremediation strategies for the cleaning of contaminated agricultural soils. Though metabolic modeling is an effective approach, the limitation of existing knowledge on the genome sequences and experimental-based biochemical data restrict its full use in various fields. Investigations on the functional properties of the microbial groups can help to improve the performance of metabolic modeling through incorporation of “omic” technologies.

Acknowledgments

Senior author is thankful to the Agricultural Research Organization, Israel for awarding the research fellowship.

References

- Adams, G. O., Fufeyin, P. T., Okoro, S. E., & Ehinomen, I. (2015). Bioremediation, biostimulation and bioaugmentation: A review. *International Journal of Environmental Bioremediation & Biodegradation*.
- Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., & Nielsen, J. (2013). The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Computational Biology*, 9(3). Available from <https://doi.org/10.1371/journal.pcbi.1002980>.
- Ahmad, K. S. (2020). Environmental contaminant 2-chloro-N-(2,6-diethylphenyl)-N-(methoxymethyl)acetamide remediation via *Xanthomonas axonopodis* and *Aspergillus niger*. *Environmental Research*, 182, 109117. Available from <https://doi.org/10.1016/j.envres.2020.109117>.
- Aislabie, J., Bej, A. K., Ryburn, J., Lloyd, N., & Wilkins, A. (2005). Characterization of *Arthrobacter nicotinovorans* HIM, an atrazine-degrading bacterium, from agricultural soil New Zealand. *FEMS Microbiology Ecology*, 52(2), 279–286. Available from <https://doi.org/10.1016/j.femsec.2004.11.012>.
- Albers, C. N., Banta, G. T., Hansen, P. E., & Jacobsen, O. S. (2008). Effect of different humic substances on the fate of diuron and its main metabolite 3,4-dichloroaniline in soil. *Environmental Science and Technology*, 42(23), 8687–8691. Available from <https://doi.org/10.1021/es800629m>.
- Ali, N., Khan, S., Li, Y., Zheng, N., & Yao, H. (2019). Influence of biochars on the accessibility of organochlorine pesticides and microbial community in contaminated soils. *Science of the Total Environment*, 647, 551–560. Available from <https://doi.org/10.1016/j.scitotenv.2018.07.425>.
- An, X., Cheng, Y., Miao, L., Chen, X., Zang, H., & Li, C. (2020). Characterization and genome functional analysis of an efficient nitrile-degrading bacterium, *Rhodococcus rhodochrous* BX2, to lay the foundation for potential bioaugmentation for remediation of nitrile-contaminated environments. *Journal of Hazardous Materials*, 389, 121906. Available from <https://doi.org/10.1016/j.jhazmat.2019.121906>.
- Annett, R., Habibi, H. R., & Hontela, A. (2014). Impact of glyphosate and glyphosate-based herbicides on the freshwater environment. *Journal of Applied Toxicology*, 34(5), 458–479. Available from <https://doi.org/10.1002/jat.2997>.
- Arfarita, N., Imai, T., Kanno, A., Yarimizu, T., Xiaofeng, S., Jie, W., ... Akada, R. (2013). The potential use of *Trichoderma viride* strain FRP3 in biodegradation of the herbicide glyphosate. *Biotechnology and Biotechnological Equipment*, 27(1), 3518–3521. Available from <https://doi.org/10.5504/bbeq.2012.0118>.
- Arora, P. K. (2015). Bacterial degradation of monocyclic aromatic amine. *Frontiers in Microbiology*, 6, 1–14. Available from <https://doi.org/10.3389/fmicb.2015.00820>.
- Ates, Ö., Oner, E. T., & Arga, K. Y. (2011). Genome-scale reconstruction of metabolic network for a halophilic extremophile, *Chromohalobacter salexigens* DSM 3043. *BMC Systems Biology*, 5, 1–13. Available from <https://doi.org/10.1186/1752-0509-5-12>.
- Badawi, N., Rønhede, S., Olsson, S., Kragelund, B. B., Johnsen, A. H., Jacobsen, O. S., & Aamand, J. (2009). Metabolites of the phenylurea herbicides chlorotoluron, diuron, isoproturon and linuron produced by the soil fungus *Mortierella* sp. *Environmental Pollution*, 157(10), 2806–2812. Available from <https://doi.org/10.1016/j.envpol.2009.04.019>.
- Bailey-serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D., & Schroeder, J. I. (2019). Genetic strategies for improving crop yields. *Nature*, 575, 109–118. Available from <https://doi.org/10.1038/s41586-019-1679-0>.
- Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B., & Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nature Protocols*. Available from <https://doi.org/10.1038/nprot.2007.99>.
- Bernat, P., Nykiel-Szymańska, J., Stolarek, P., Słaba, M., Szewczyk, R., & Różalska, S. (2018). 2,4-Dichlorophenoxyacetic acid-induced oxidative stress: Metabolome and membrane modifications in *umbelopsis isabellina*, a herbicide degrader. *PLoS One*. Available from <https://doi.org/10.1371/journal.pone.0199677>.
- Bers, K., Leroy, B., Breugelmans, P., Albers, P., Lavigne, R., Sørensen, S. R., ... Springael, D. (2011). A novel hydrolase identified by genomic-proteomic analysis of phenylurea herbicide mineralization by *Variovorax* sp. strain SRS16. *Applied and Environmental Microbiology*, 77(24), 8754–8764. Available from <https://doi.org/10.1128/AEM.06162-11>.
- Bezerra, C., de, O., Cusioli, L. F., Quesada, H. B., Nishi, L., Mantovani, D., ... Bergamasco, R. (2020). Assessment of the use of *Moringa oleifera* seed husks for removal of pesticide diuron from contaminated water. *Environmental Technology (United Kingdom)*, 41(2), 191–201. Available from <https://doi.org/10.1080/09593330.2018.1493148>.
- Biggs, M. B., Medlock, G. L., Kolling, G. L., & Papin, J. A. (2015). Metabolic network modeling of microbial communities. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(5), 317–334. Available from <https://doi.org/10.1002/wsbm.1308>.

- Billet, L., Devers-Lamrani, M., Serre, R.-F., Julia, E., Vandecasteele, C., Rouard, N., ... Spor, A. (2021). Complete genome sequences of four atrazine-degrading bacterial strains, *Pseudomonas* sp. strain ADPe, *Arthrobacter* sp. strain TES, *Variovorax* sp. strain 38R, and *Chelatobacter* sp. strain SR38. *Microbiology Resource Announcements*. Available from <https://doi.org/10.1128/mra.01080-20>.
- Boundy-Mills, K. L., De Souza, M. L., Mandelbaum, R. T., Wackett, L. P., & Sadowsky, M. J. (1997). The *atzB* gene of *Pseudomonas* sp. strain ADP encodes the second enzyme of a novel atrazine degradation pathway. *Applied and Environmental Microbiology*, 63(3), 916–923. Available from <https://doi.org/10.1128/aem.63.3.916-923.1997>.
- Briceño, G., Fuentes, M. S., Saez, J. M., Diez, M. C., & Benimeli, C. S. (2018). *Streptomyces* genus as biotechnological tool for pesticide degradation in polluted systems. *Critical Reviews in Environmental Science and Technology*, 48(10–12), 773–805. Available from <https://doi.org/10.1080/10643389.2018.1476958>.
- Briceño, G., Palma, G., & Durán, N. (2007). Influence of organic amendment on the biodegradation and movement of pesticides. *Critical Reviews in Environmental Science and Technology*, 37(3). Available from <https://doi.org/10.1080/10643380600987406>.
- Carles, L., Rossi, F., Besse-Hoggan, P., Blavignac, C., Leremboure, M., Artigas, J., & Batisson, I. (2018). Nicosulfuron degradation by an ascomycete fungus isolated from submerged alnus leaf litter. *Frontiers in Microbiology*. Available from <https://doi.org/10.3389/fmicb.2018.03167>.
- Carranza, C. S., Regñicoli, J. P., Aluffi, M. E., Benito, N., Chiachiera, S. M., Barberis, C. L., & Magnoli, C. E. (2019). Glyphosate in vitro removal and tolerance by *Aspergillus oryzae* in soil microcosms. *International Journal of Environmental Science and Technology*, 16(12), 7673–7682. Available from <https://doi.org/10.1007/s13762-019-02347-x>.
- Chan, C. Y., Chan, H. S., & Wong, P. K. (2019). Integrated photocatalytic-biological treatment of triazine-containing pollutants. *Chemosphere*, 222, 371–380. Available from <https://doi.org/10.1016/j.chemosphere.2019.01.127>.
- Coelho-Moreira, J., da, S., Brugnari, T., Sá-Nakanishi, A. B., Castoldi, R., de Souza, C. G. M., ... Peralta, R. M. (2018). Evaluation of diuron tolerance and biotransformation by the white-rot fungus *Ganoderma lucidum*. *Fungal Biology*, 122(6), 471–478. Available from <https://doi.org/10.1016/j.funbio.2017.10.008>.
- Cycoń, M., Mroziak, A., & Piotrowska-Seget, Z. (2017). Bioaugmentation as a strategy for the remediation of pesticide-polluted soil: A review. *Chemosphere*. Available from <https://doi.org/10.1016/j.chemosphere.2016.12.129>.
- D506-D515. (2019). UniProt: A worldwide hub of protein knowledge The UniProt Consortium. *Nucleic Acids Research*. Available from <https://doi.org/10.1093/nar/gky1049>.
- Dai, Y., Li, N., Zhao, Q., & Xie, S. (2015). Bioremediation using *Novosphingobium* strain DY4 for 2,4-dichlorophenoxyacetic acid-contaminated soil and impact on microbial community structure. *Biodegradation*, 26(2), 161–170. Available from <https://doi.org/10.1007/s10532-015-9724-7>.
- Delgado-Moreno, L., & Peña, A. (2007). Organic amendments from olive cake as a strategy to modify the degradation of sulfonylurea herbicides in soil. *Journal of Agricultural and Food Chemistry*, 55(15), 6213–6218. Available from <https://doi.org/10.1021/jf0708342>.
- Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A. A., & Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and model SEED. *Methods in Molecular Biology*. Available from https://doi.org/10.1007/978-1-62703-299-5_2.
- Dhiman, N., Jasrotia, T., Sharma, P., Negi, S., Chaudhary, S., Kumar, R., ... Kumar, R. (2020). Immobilization interaction between xenobiotic and *Bjerkandera adusta* for the biodegradation of atrazine. *Chemosphere*. Available from <https://doi.org/10.1016/j.chemosphere.2020.127060>.
- Duke, S. O. (2020). Glyphosate: Environmental fate and impact. *Weed Science*, 68(3), 201–207. Available from <https://doi.org/10.1017/wsc.2019.28>.
- Dwivedi, S., Singh, B. R., Al-Khedhairi, A. A., & Musarrat, J. (2011). Biodegradation of isoproturon using a novel *Pseudomonas aeruginosa* strain JS-11 as a multi-functional bioinoculant of environmental significance. *Journal of Hazardous Materials*, 185(2–3), 938–944. Available from <https://doi.org/10.1016/j.jhazmat.2010.09.110>.
- Egea, T. C., da Silva, R., Boscolo, M., Rigonato, J., Monteiro, D. A., Grünig, D., ... Gomes, E. (2017). Diuron degradation by bacteria from soil of sugarcane crops. *Heliyon*, 3(12), e00471. Available from <https://doi.org/10.1016/j.heliyon.2017.e00471>.
- Eibisch, N., Schroll, R., & Fuß, R. (2015). Effect of pyrochar and hydrochar amendments on the mineralization of the herbicide isoproturon in an agricultural soil. *Chemosphere*, 134, 528–535. Available from <https://doi.org/10.1016/j.chemosphere.2014.11.074>.
- El-Deeb, B. A., Soltan, S. M., Ali, A. M., & Ali, K. A. (2000). Detoxication of the herbicide Diuron by *pseudomonas* sp. *Folia Microbiologica*, 45(3), 211–216. Available from <https://doi.org/10.1007/bf02908946>.
- Elgueta, S., Santos, C., Lima, N., & Diez, M. C. (2016). Immobilization of the white-rot fungus *Anthracoxyllum discolor* to degrade the herbicide atrazine. *AMB Express*. Available from <https://doi.org/10.1186/s13568-016-0275-z>.
- Esparza-Naranjo, S. B., da Silva, G. F., Duque-Castaño, D. C., Araújo, W. L., Peres, C. K., Boroski, M., & Bonugli-Santos, R. C. (2020). Potential for the biodegradation of atrazine using leaf litter fungi from a subtropical protection area. *Current Microbiology*, 78(1), 358–368. Available from <https://doi.org/10.1007/s00284-020-02288-6>.
- Esquirol, L., Peat, T. S., Wilding, M., Hartley, C. J., Newman, J., & Scott, C. (2018). A novel decarboxylating amidohydrolase involved in avoiding metabolic dead ends during cyanuric acid catabolism in *Pseudomonas* sp. Strain ADP. *PLoS One*, 13(11), 1–17. Available from <https://doi.org/10.1371/journal.pone.0206949>.
- Faria, J. P., Rocha, M., Rocha, I., & Henry, C. S. (2018). Methods for automated genome-scale metabolic model reconstruction. *Biochemical Society Transactions*, 46(4), 931–936. Available from <https://doi.org/10.1042/BST20170246>.
- Faust, K. (2019). Microbial consortium design benefits from metabolic modeling. *Trends in Biotechnology*. Available from <https://doi.org/10.1016/j.tibtech.2018.11.004>.
- Feist, A. M., Scholten, J. C. M., Palsson, B., Brockman, F. J., & Ideker, T. (2006). Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Molecular Systems Biology*, 2, 1–14. Available from <https://doi.org/10.1038/msb4100046>.

- Finley, S. D., Broadbelt, L. J., & Hatzimanikatis, V. (2009). Computational framework for predictive biodegradation. *Biotechnology and Bioengineering*, 104(6), 1086–1097. Available from <https://doi.org/10.1002/bit.22489>.
- Firdous, S., Iqbal, S., & Anwar, S. (2020). Optimization and modeling of glyphosate biodegradation by a novel *Comamonas odontotermitis* P2 through response surface methodology. *Pedosphere*. Available from [https://doi.org/10.1016/S1002-0160\(17\)60381-3](https://doi.org/10.1016/S1002-0160(17)60381-3).
- Flahaut, N. A. L., Wiersma, A., Van De Bunt, B., Martens, D. E., Schaap, P. J., Sijtsma, L., . . . De Vos, W. M. (2013). Genome-scale metabolic model for *Lactococcus lactis* MG1363 and its application to the analysis of flavor formation. *Applied Microbiology and Biotechnology*, 97(19), 8729–8739. Available from <https://doi.org/10.1007/s00253-013-5140-2>.
- Fondi, M., Maida, I., Perrin, E., Mellera, A., Mocali, S., Parrilli, E., . . . Fani, R. (2015). Genome-scale metabolic reconstruction and constraint-based modelling of the Antarctic bacterium *Pseudoalteromonas haloplanktis*TAC125. *Environmental Microbiology*, 17(3), 751–766. Available from <https://doi.org/10.1111/1462-2920.12513>.
- Forouzanoghar, M., Haghnia, G. H., & Koocheki, A. (2005). Organic amendments to enhance atrazine and metolachlor degradation in two contaminated soils with contrasting textures. *Soil and Sediment Contamination*, 14(4), 345–355. Available from <https://doi.org/10.1080/15320380590954060>.
- Forouzesh, A., Zand, E., Soufizadeh, S., & Samadi Foroushani, S. (2015). Classification of herbicides according to chemical family for weed resistance management strategies—an update. *Weed Research*, 55(4), 334–358. Available from <https://doi.org/10.1111/wre.12153>.
- Gámiz, B., Celis, R., Hermosín, M. C., & Cornejo, J. (2010). Organoclays as soil amendments to increase the efficacy and reduce the environmental impact of the herbicide fluometuron in agricultural soils. *Journal of Agricultural and Food Chemistry*. Available from <https://doi.org/10.1021/jf100760s>.
- García-Romero, I., Nogales, J., Díaz, E., Santero, E., & Floriano, B. (2020). Understanding the metabolism of the tetralin degrader *Sphingopyxis granulii* strain TFA through genome-scale metabolic modelling. *Scientific Reports*, 10(1), 1–14. Available from <https://doi.org/10.1038/s41598-020-65258-9>.
- Geed, S. R., Prasad, S., Kureel, M. K., Singh, R. S., & Rai, B. N. (2018). Biodegradation of wastewater in alternating aerobic-anoxic lab scale pilot plant by *Alcaligenes* sp. S3 isolated from agricultural field. *Journal of Environmental Management*, 214, 408–415. Available from <https://doi.org/10.1016/j.jenvman.2018.03.031>.
- Ghatge, S., Yang, Y., Moon, S., Song, W. Y., Kim, T. Y., Liu, K. H., & Hur, H. G. (2021). A novel pathway for initial biotransformation of dinitroaniline herbicide butralin from a newly isolated bacterium *Sphingopyxis* sp. strain HMH. *Journal of Hazardous Materials*. Available from <https://doi.org/10.1016/j.jhazmat.2020.123510>.
- Giacomazzi, S., & Cochet, N. (2004). Environmental impact of diuron transformation: A review. *Chemosphere*, 56(11), 1021–1032. Available from <https://doi.org/10.1016/j.chemosphere.2004.04.061>.
- Gong, X., Huang, D., Liu, Y., Peng, Z., Zeng, G., Xu, P., . . . Wan, J. (2018). Remediation of contaminated soils by biotechnology with nanomaterials: Bio-behavior, applications, and perspectives. *Critical Reviews in Biotechnology*, 38(3), 455–468. Available from <https://doi.org/10.1080/07388551.2017.1368446>.
- Goyal, N., Widiastuti, H., Karimi, I. A., & Zhou, Z. (2014). A genome-scale metabolic model of *Methanococcus maripaludis* S2 for CO₂ capture and conversion to methane. *Molecular Biosystems*, 10(5), 1043–1054. Available from <https://doi.org/10.1039/c3mb70421a>.
- Grenni, P., Caracciolo, A. B., Rodríguez-Cruz, M. S., & Sánchez-Martín, M. J. (2009). Changes in the microbial activity in a soil amended with oak and pine residues and treated with linuron herbicide. *Applied Soil Ecology*, 41(1), 2–7. Available from <https://doi.org/10.1016/j.apsoil.2008.07.006>.
- Grenni, P., Rodríguez-Cruz, M. S., Herrero-Hernández, E., Marín-Benito, J. M., Sánchez-Martín, M. J., & Caracciolo, A. B. (2012). Effects of wood amendments on the degradation of terbutylazine and on soil microbial community activity in a clay loam soil. *Water, Air, and Soil Pollution*, 223(8), 5401–5412. Available from <https://doi.org/10.1007/s11270-012-1289-z>.
- Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., . . . Dubchak, I. (2012). The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research*. Available from <https://doi.org/10.1093/nar/gkr947>.
- Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, 20(1), 1–18. Available from <https://doi.org/10.1186/s13059-019-1730-3>.
- Guimarães, A. C. D., Mendes, K. F., dos Reis, F. C., Campion, T. F., Christoffoleti, P. J., & Tornisielo, V. L. (2018). Role of soil physicochemical properties in quantifying the fate of diuron, hexazinone, and metribuzin. *Environmental Science and Pollution Research*. Available from <https://doi.org/10.1007/s11356-018-1469-5>.
- Hatakeyama, T., Takagi, K., Yamazaki, K., Sakakibara, F., Ito, K., Takasu, E., . . . Fujii, K. (2015). Mineralization of melamine and cyanuric acid as sole nitrogen source by newly isolated *Arthrobacter* spp. using a soil-charcoal perfusion method. *World Journal of Microbiology and Biotechnology*. Available from <https://doi.org/10.1007/s11274-015-1832-3>.
- Haynes, D., Ralph, P., Prange, J., & Dennison, B. (2000). The impact of the herbicide diuron on photosynthesis in three species of tropical seagrass. *Marine Pollution Bulletin*, 41, 288–293.
- He, H., Liu, Y., You, S., Liu, J., Xiao, H., & Tu, Z. (2019). A review on recent treatment technology for herbicide atrazine in contaminated environment. *International Journal of Environmental Research and Public Health*, 16(24). Available from <https://doi.org/10.3390/ijerph16245129>.
- Henry, C. S., Bernstein, H. C., Weisenhorn, P., Taylor, R. C., Lee, J. Y., Zucker, J., & Song, H. S. (2016). Microbial community metabolic modeling: A community data-driven network reconstruction. *Journal of Cellular Physiology*. Available from <https://doi.org/10.1002/jcp.25428>.
- Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9), 977–982. Available from <https://doi.org/10.1038/nbt.1672>.

- Henry, C. S., Zinner, J. F., Cohoon, M. P., & Stevens, R. L. (2009). iBsu1103: A new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biology*, 10(6), 1–15. Available from <https://doi.org/10.1186/gb-2009-10-6-r69>.
- Henson, M. A., & Hanly, T. J. (2014). Dynamic flux balance analysis for synthetic microbial communities. *IET Systems Biology*. Available from <https://doi.org/10.1049/iet-syb.2013.0021>.
- Hinteregger, C., Loidl, M., & Streichsbier, F. (1992). Characterization of isofunctional ring-cleaving enzymes in aniline and 3-chloroaniline degradation by *Pseudomonas acidovorans* CA28. *FEMS Microbiology Letters*, 97(3), 261–266. Available from [https://doi.org/10.1016/0378-1097\(92\)90346-P](https://doi.org/10.1016/0378-1097(92)90346-P).
- Hongsawat, P., & Vangnai, A. S. (2011). Biodegradation pathways of chloroanilines by *Acinetobacter baylyi* strain GFJ2. *Journal of Hazardous Materials*, 186(2–3), 1300–1307. Available from <https://doi.org/10.1016/j.jhazmat.2010.12.002>.
- Hou, Y., Dong, W., Wang, F., Li, J., Shen, W., Li, Y., & Cui, Z. (2014). Degradation of acetochlor by a bacterial consortium of *Rhodococcus* sp. T3-1, *Delftia* sp. T3-6 and *Sphingobium* sp. MEA3-1. *Letters in Applied Microbiology*. Available from <https://doi.org/10.1111/lam.12242>.
- Hussain, S., Arshad, M., Springael, D., Sørensen, S. R., Bending, G. D., Devers-Lamrani, M., . . . Martin-Laurent, F. (2015). Abiotic and biotic processes governing the fate of Phenylurea herbicides in soils: A review. *Critical Reviews in Environmental Science and Technology*. Available from <https://doi.org/10.1080/10643389.2014.1001141>.
- Hussain, S., Sørensen, S. R., Devers-Lamrani, M., El-Sebai, T., & Martin-Laurent, F. (2009). Characterization of an isotopuron mineralizing bacterial culture enriched from a French agricultural soil. *Chemosphere*, 77(8), 1052–1059. Available from <https://doi.org/10.1016/j.chemosphere.2009.09.020>.
- Islam, F., Wang, J., Farooq, M. A., Khan, M. S. S., Xu, L., Zhu, J., . . . Zhou, W. (2018). Potential impact of the herbicide 2,4-dichlorophenoxyacetic acid on human and ecosystems. *Environment International*, 111, 332–351. Available from <https://doi.org/10.1016/j.envint.2017.10.020>.
- Islam, M. M., Le, T., Daggumati, S. R., & Saha, R. (2020). Investigation of microbial community interactions between lake Washington methanotrophs using genome-scale metabolic modeling. *BioRxiv*. Available from <https://doi.org/10.1101/2020.02.21.958074>.
- Jablonski, N. D., Schäffer, A., & Burauel, P. (2011). Still present after all these years: Persistence plus potential toxicity raise questions about the use of atrazine. *Environmental Science and Pollution Research*, 18(2), 328–331. Available from <https://doi.org/10.1007/s11356-010-0431-y>.
- Jo, J., Oh, J., & Park, C. (2020). Microbial community analysis using high-throughput sequencing technology: A beginner's guide for microbiologists. *Journal of Microbiology*, 58(3), 176–192. Available from <https://doi.org/10.1007/s12275-020-9525-5>.
- Kadian, N., Gupta, A., Satya, S., Mehta, R. K., & Malik, A. (2008). Biodegradation of herbicide (atrazine) in contaminated soil using various bioprocessed materials. *Bioresource Technology*, 99(11), 4642–4647. Available from <https://doi.org/10.1016/j.biortech.2007.06.064>.
- Kah, M. (2020). Emerging investigator series: Nanotechnology to develop novel agrochemicals : Critical issues to consider in the global agricultural context. *Environmental Science Nano*, 1867–1873. Available from <https://doi.org/10.1039/d0en00271b>.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1), D457–D462. Available from <https://doi.org/10.1093/nar/gkv1070>.
- Kanissery, R. G., & Sims, G. K. (2011). Biostimulation for the enhanced degradation of herbicides in soil. *Applied and Environmental Soil Science*. Available from <https://doi.org/10.1155/2011/843450>.
- Kaur, R., & Goyal, D. (2020). Biodegradation of butachlor by *Bacillus altitudinis* and identification of metabolites. *Current Microbiology*. Available from <https://doi.org/10.1007/s00284-020-02031-1>.
- Khatoon, H., & Rai, J. P. N. (2020). Optimization studies on biodegradation of atrazine by *Bacillus badius* ABP6 strain using response surface methodology. *Biotechnology Reports*, 26, e00459. Available from <https://doi.org/10.1016/j.btre.2020.e00459>.
- Kim, Y. M., Park, K., Kim, W. C., Shin, J. H., Kim, J. E., Park, H. D., & Rhee, I. K. (2007). Cloning and characterization of a catechol-degrading gene cluster from 3,4-dichloroaniline degrading bacterium *Pseudomonas* sp. KB35B. *Journal of Agricultural and Food Chemistry*, 55(12), 4722–4727. Available from <https://doi.org/10.1021/jf070116f>.
- Kniss, A. R., & Coburn, C. W. (2015). Quantitative evaluation of the environmental impact quotient (EIQ) for comparing herbicides. *PLoS One*, 10(6), 1–13. Available from <https://doi.org/10.1371/journal.pone.0131200>.
- Koch, S., Kohrs, F., Lahmann, P., Bissinger, T., Wendschuh, S., Benndorf, D., . . . Klamt, S. (2019). Redcom: A strategy for reduced metabolic modeling of complex microbial communities and its application for analyzing experimental datasets from anaerobic digestion. *PLoS Computational Biology*, 15(2). Available from <https://doi.org/10.1371/journal.pcbi.1006759>.
- Kovács, K., Farkas, J., Veréb, G., Arany, E., Simon, G., Schrantz, K., . . . Alapi, T. (2016). Comparison of various advanced oxidation processes for the degradation of phenylurea herbicides. *Journal of Environmental Science and Health - Part B Pesticides, Food Contaminants, and Agricultural Wastes*, 51(4), 205–214. Available from <https://doi.org/10.1080/03601234.2015.1120597>.
- Kulyashov, M., Peltek, S. E., & Akberdin, I. R. (2020). A genome-scale metabolic model of 2,3-butanediol production by thermophilic bacteria *Geobacillus icigianus*. *Microorganisms*, 8(7), 1–13. Available from <https://doi.org/10.3390/microorganisms8071002>.
- Lachance, J. C., Lloyd, C. J., Monk, J. M., Yang, L., Sastry, A. V., Seif, Y., . . . Jacques, P. É. (2019). BOFDAT: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS Computational Biology*. Available from <https://doi.org/10.1371/journal.pcbi.1006971>.
- Li, Y., Rashid, A., Wang, H., Hu, A., Lin, L., Yu, C. P., . . . Sun, Q. (2018). Contribution of biotic and abiotic factors in the natural attenuation of sulfamethoxazole: A path analysis approach. *Science of the Total Environment*, 633, 1217–1226. Available from <https://doi.org/10.1016/j.scitotenv.2018.03.232>.
- Lieven, C., Petersen, L., Jørgensen, S. B., Gernaey, K., Herrgard, M., & Sonnenschein, N. (2018). A genome-scale metabolic model for *Methylococcus capsulatus* predicts reduced efficiency uphill electron transfer to pMMO. *BioRxiv*. Available from <https://doi.org/10.1101/329714>.

- Lin, Z., Zhen, Z., Ren, L., Yang, J., Luo, C., Zhong, L., ... Zhang, D. (2018). Effects of two ecological earthworm species on atrazine degradation performance and bacterial community structure in red soil. *Chemosphere*, 196, 467–475. Available from <https://doi.org/10.1016/j.chemosphere.2017.12.177>.
- Liu, J., Hua, R., Lv, P., Tang, J., Wang, Y., Cao, H., ... Li, Q. X. (2017). Novel hydrolytic de-methylthiolation of the s-triazine herbicide prometryn by *Leucobacter* sp. JW-1. *Science of the Total Environment*. Available from <https://doi.org/10.1016/j.scitotenv.2016.11.006>.
- Liu, Y., Lonappan, L., Brar, S. K., & Yang, S. (2018). Impact of biochar amendment in agricultural soils on the sorption, desorption, and degradation of pesticides: A review. *Science of the Total Environment*, 645, 60–70. Available from <https://doi.org/10.1016/j.scitotenv.2018.07.099>.
- Lopes, R., de, O., Pereira, P. M., Pereira, A. R. B., Fernandes, K. V., Carvalho, J. F., ... Ferreira-Leitão, V. S. (2020). Atrazine, desethylatrazine (DEA) and desisopropylatrazine (DIA) degradation by *Pleurotus ostreatus* INCQS 40310. *Biocatalysis and Biotransformation*. Available from <https://doi.org/10.1080/10242422.2020.1754805>.
- Madureira Barroso, G., dos Santos, J. B., de Oliveira, I. T., Rocha Nunes, T. K. M., Alves Ferreira, E., Marinho Pereira, I., ... de Freitas Souza, M. (2020). Tolerance of *Bradyrhizobium* sp. BR 3901 to herbicides and their ability to use these pesticides as a nutritional source. *Ecological Indicators*. Available from <https://doi.org/10.1016/j.ecolind.2020.106783>.
- Mahadevan, R., & Henson, M. A. (2012). Genome-based modeling and design of metabolic interactions in microbial communities. *Computational and Structural Biotechnology Journal*, 3(4), e201210008. Available from <https://doi.org/10.5936/cs bj.201210008>.
- Marín-Benito, J. M., Barba, V., Ordax, J. M., Sánchez-Martín, M. J., & Rodríguez-Cruz, M. S. (2018). Recycling organic residues in soils as amendments: Effect on the mobility of two herbicides under different management practices. *Journal of Environmental Management*, 224, 172–181. Available from <https://doi.org/10.1016/j.jenvman.2018.07.045>.
- Marín-Benito, J. M., Carpio, M. J., Sánchez-Martín, M. J., & Rodríguez-Cruz, M. S. (2019). Previous degradation study of two herbicides to simulate their fate in a sandy loam soil: Effect of the temperature and the organic amendments. *Science of the Total Environment*, 653, 1301–1310. Available from <https://doi.org/10.1016/j.scitotenv.2018.11.015>.
- Martins, M., Dairou, J., Rodrigues-Lima, F., Dupret, J. M., & Silar, P. (2010). Insights into the phylogeny of arylamine N-acetyltransferases in fungi. *Journal of Molecular Evolution*. Available from <https://doi.org/10.1007/s00239-010-9371-x>.
- Meena, R. S., Kumar, S., Datta, R., Lal, R., Vijayakumar, V., Brtnicky, M., ... Marfo, D. (2020). *Impact of Agrochemicals on Soil Microbiota and Management: A Review*. *Land* 9, 2: 34.
- Mehdizadeh, M., Izadi-Darbandi, E., Naseri Pour Yazdi, M. T., Rastgoo, M., Malaekheh-Nikouei, B., & Nassirli, H. (2019). Impacts of different organic amendments on soil degradation and phytotoxicity of metribuzin. *International Journal of Recycling of Organic Waste in Agriculture*, 8 (s1), 113–121. Available from <https://doi.org/10.1007/s40093-019-0280-8>.
- Mendoza, S. N., Olivier, B. G., Molenaar, D., & Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biology*, 20(1), 1–20. Available from <https://doi.org/10.1186/s13059-019-1769-1>.
- Mohamed, A. T., El Hussein, A. A., El Siddig, M. A., & Osman, A. G. (2011). Degradation of oxyfluorfen herbicide by soil microorganisms biodegradation of herbicides. *Biotechnology (Reading, Mass.)*, 10(3), 274–279. Available from <https://doi.org/10.3923/biotech.2011.274.279>.
- Montagud, A., Navarro, E., Fernández de Córdoba, P., Urchueguía, J. F., & Patil, K. R. (2010). Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Systems Biology*, 4. Available from <https://doi.org/10.1186/1752-0509-4-156>.
- Moorman, T. B., Cowan, J. K., Arthur, E. L., & Coats, J. R. (2001). Organic amendments to enhance herbicide biodegradation in contaminated soils. *Biology and Fertility of Soils*, 33(6), 541–545. Available from <https://doi.org/10.1007/s003740100367>.
- Morán, A. C., Müller, A., Manzano, M., & González, B. (2006). Simazine treatment history determines a significant herbicide degradation potential in soils that is not improved by bioaugmentation with *Pseudomonas* sp. ADP. *Journal of Applied Microbiology*, 101(1), 26–35. Available from <https://doi.org/10.1111/j.1365-2672.2006.02990.x>.
- Morgante, V., López-López, A., Flores, C., González, M., González, B., Vázquez, M., ... Seeger, M. (2010). Bioaugmentation with *Pseudomonas* sp. strain MHP41 promotes simazine attenuation and bacterial community changes in agricultural soils. *FEMS Microbiology Ecology*, 71(1), 114–126. Available from <https://doi.org/10.1111/j.1574-6941.2009.00790.x>.
- Mudhoo, A., & Garg, V. K. (2011). Sorption, transport and transformation of atrazine in soils, minerals and composts: A review. *Pedosphere*. Available from [https://doi.org/10.1016/S1002-0160\(10\)60074-4](https://doi.org/10.1016/S1002-0160(10)60074-4).
- Mueller, T. C., Parker, E. T., Steckel, L., Clay, S. A., Owen, M. D. K., Curran, W. S., ... Klein, R. (2017). Enhanced atrazine degradation is widespread across the United States. *Pest Management Science*, 73(9), 1953–1961. Available from <https://doi.org/10.1002/ps.4566>.
- Mukherjee, I. (2009). Effect of organic Amendments on degradation of Atrazine. *Bulletin of Environmental Contamination and Toxicology*, 83(6), 832–835. Available from <https://doi.org/10.1007/s00128-009-9849-7>.
- Nagarajan, H., Sahin, M., Nogales, J., Latif, H., Lovley, D. R., Ebrahim, A., & Zengler, K. (2013). Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of *Clostridium ljungdahlii*. *Microbial Cell Factories*, 12(1), 1–13. Available from <https://doi.org/10.1186/1475-2859-12-118>.
- Nanda, P., Patra, P., Das, M., & Ghosh, A. (2020). Reconstruction and analysis of genome-scale metabolic model of weak Crabtree positive yeast *Lachancea kluyveri*. *Scientific Reports*, 10(1), 1–18. Available from <https://doi.org/10.1038/s41598-020-73253-3>.
- Nogales, J., Palsson, B., & Thiele, I. (2008). A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Systems Biology*, 2, 1–20. Available from <https://doi.org/10.1186/1752-0509-2-79>.
- Nykiel-Szymańska, J., Bernat, P., & Słaba, M. (2018). Potential of *Trichoderma koningii* to eliminate alachlor in the presence of copper ions. *Ecotoxicology and Environmental Safety*. Available from <https://doi.org/10.1016/j.ecoenv.2018.06.060>.

- Ofaim, S., Zarecki, R., Porob, S., Gat, D., Lahav, T., Kashi, Y., ... Freilich, S. (2020). Genome-scale reconstruction of *Paenarthrobacter aurescens* TC1 metabolic model towards the study of atrazine bioremediation. *Scientific Reports*, 10(1), 13019. Available from <https://doi.org/10.1038/s41598-020-69509-7>.
- Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., ... Stevens, R. (2014). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Research*, 42(D1), 206–214. Available from <https://doi.org/10.1093/nar/gkt1226>.
- Öztürk, B., Werner, J., Meier-Kolthoff, J. P., Bunk, B., Spröer, C., & Springael, D. (2020). Comparative genomics suggests mechanisms of genetic adaptation toward the catabolism of the phenylurea herbicide linuron in *variovorax*. *Genome Biology and Evolution*, 12(6), 827–841. Available from <https://doi.org/10.1093/gbe/evaa085>.
- Pacwa-Plociniczak, M., Plaza, G. A., & Piotrowska-Seget, Z. (2016). Monitoring the changes in a bacterial community in petroleum-polluted soil bioaugmented with hydrocarbon-degrading strains. *Applied Soil Ecology*, 105, 76–85. Available from <https://doi.org/10.1016/j.apsoil.2016.04.005>.
- Pan, X., Wang, S., Shi, N., Fang, H., & Yu, Y. (2018). Biodegradation and detoxification of chlorimuron-ethyl by *Enterobacter ludwigii* sp. CE-1. *Ecotoxicology and Environmental Safety*. Available from <https://doi.org/10.1016/j.ecoenv.2017.12.023>.
- Perissini-Lopes, B., Egea, T. C., Monteiro, D. A., Vici, A. C., Da Silva, D. G. H., Lisboa, D. C. D. O., ... Gomes, E. (2016). Evaluation of diuron tolerance and biotransformation by fungi from a sugar cane plantation sandy-loam soil. *Journal of Agricultural and Food Chemistry*, 64(49), 9268–9275. Available from <https://doi.org/10.1021/acs.jafc.6b03247>.
- Peterson, M. A., McMaster, S. A., Riechers, D. E., Skelton, J., & Stahlman, P. W. (2016). 2,4-D past, present, and future: A review. *Weed Technology*, 30(2), 303–345. Available from <https://doi.org/10.1614/wt-d-15-00131.1>.
- Pileggi, M., Pileggi, S. A. V., & Sadowsky, M. J. (2020). Herbicide bioremediation: From strains to bacterial communities. *Heliyon*, 6(12). Available from <https://doi.org/10.1016/j.heliyon.2020.e05767>.
- Piubeli, F., Salvador, M., Argandoña, M., Nieto, J. J., Bernal, V., Pastor, J. M., ... Vargas, C. (2018). Insights into metabolic osmoadaptation of the ectoines-producer bacterium *Chromohalobacter salexigens* through a high-quality genome scale metabolic model. *Microbial Cell Factories*, 17(1), 1–20. Available from <https://doi.org/10.1186/s12934-017-0852-0>.
- Piutti, S., Marchand, A. L., Lagacherie, B., Martin-Laurent, F., & Soulas, G. (2002). Effect of cropping cycles and repeated herbicide applications on the degradation of diclofopmethyl, bentazone, diuron, isoproturon and pendimethalin in soil. *Pest Management Science*. Available from <https://doi.org/10.1002/ps.459>.
- Pose-Juan, E., Sánchez-Martín, M. J., Herrero-Hernández, E., & Rodríguez-Cruz, M. S. (2015). Application of mesotrione at different doses in an amended soil: Dissipation and effect on the soil microbial biomass and activity. *Science of the Total Environment*, 536, 31–38. Available from <https://doi.org/10.1016/j.scitotenv.2015.07.039>.
- Puchałka, J., Oberhardt, M. A., Godinho, M., Bielecka, A., Regenhardt, D., Timmis, K. N., ... Martins Dos Santos, V. A. P. (2008). Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. *PLoS Computational Biology*, 4(10). Available from <https://doi.org/10.1371/journal.pcbi.1000210>.
- Qu, R., He, B., Yang, J., Lin, H., Yang, W., Wu, Q., ... Yang, G. (2021). Where are the new herbicides? *Pest Management Science*, 0–3. Available from <https://doi.org/10.1002/ps.6285>.
- Richmond, M. E. (2018). Glyphosate: A review of its global use, environmental impact, and potential health effects on humans and other species. *Journal of Environmental Studies and Sciences*, 8(4), 416–434. Available from <https://doi.org/10.1007/s13412-018-0517-2>.
- Röling, W. F. M., & Van Bodegom, P. M. (2014). Toward quantitative understanding on microbial community structure and functioning: A modeling-centered approach using degradation of marine oil spills as example. *Frontiers in Microbiology*, 5, 1–12. Available from <https://doi.org/10.3389/fmicb.2014.00125>.
- Rosculete, C. A., Bonciu, E., Rosculete, E., & Olaru, L. A. (2019). Determination of the environmental pollution potential of some herbicides by the assessment of cytotoxic and genotoxic effects on *Allium cepa*. *International Journal of Environmental Research and Public Health*, 16(1). Available from <https://doi.org/10.3390/ijerph16010075>.
- Rubio-Bellido, M., Morillo, E., & Villaverde, J. (2018). Assessment of soil diuron bioavailability to plants and microorganisms through non-exhaustive chemical extractions of the herbicide. *Geoderma*, 312, 130–138. Available from <https://doi.org/10.1016/j.geoderma.2017.09.031>.
- Safak Yilmaz, L., & Walhout, A. J. M. (2016). A *Caenorhabditis elegans* genome-scale metabolic network model. *Cell Systems*, 2(5), 297–311. Available from <https://doi.org/10.1016/j.cels.2016.04.012>.
- Sagarkar, S., Bhardwaj, P., Storck, V., Devers-Lamrani, M., Martin-Laurent, F., & Kapley, A. (2016). s-triazine degrading bacterial isolate *Arthrobacter* sp. AK-YN10, a candidate for bioaugmentation of atrazine contaminated soil. *Applied Microbiology and Biotechnology*, 100(2), 903–913. Available from <https://doi.org/10.1007/s00253-015-6975-5>.
- Salimi, F., Zhuang, K., & Mahadevan, R. (2010). Genome-scale metabolic modeling of a clostridial co-culture for consolidated bioprocessing. *Biotechnology Journal*. Available from <https://doi.org/10.1002/biot.201000159>.
- Santos, L. H. M. L. M., Freixa, A., Insa, S., Acuña, V., Sanchís, J., Farré, M., ... Rodríguez-Mozaz, S. (2019). Impact of fullerenes in the bioaccumulation and biotransformation of venlafaxine, diuron and triclosan in river biofilms. *Environmental Research*, 169, 377–386. Available from <https://doi.org/10.1016/j.envres.2018.11.036>.
- Saravanan, A., Kumar, P. S., Vo, D. V. N., Yaashikaa, P. R., Karishma, S., Jeevanantham, S., ... Bharathi, V. D. (2020). Photocatalysis for removal of environmental pollutants and fuel production: A review. *Environmental Chemistry Letters*, 0123456789. Available from <https://doi.org/10.1007/s10311-020-01077-8>.

- Schellenberger, J., Park, J. O., Conrad, T. M., & Palsson, B. T. (2010). BiGG: A biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*. Available from <https://doi.org/10.1186/1471-2105-11-213>.
- Sharma, P., & Suri, C. R. (2011). Biotransformation and biomonitoring of phenylurea herbicide diuron. *Bioresource Technology*, *102*(3), 3119–3125. Available from <https://doi.org/10.1016/j.biortech.2010.10.076>.
- Silva, T. M., Stets, M. I., Mazzetto, A. M., Andrade, D., Pileggi, F., Fávero, S. A. V., . . . Pileggi, M. (2007). Degradation of 2,4-D herbicide by microorganisms isolated from Brazilian contaminated soil. *Brazilian Journal of Microbiology*, *38*(3), 522–525. Available from <https://doi.org/10.1590/S1517-83822007000300026>.
- Singh, B., & Singh, K. (2016). Microbial degradation of herbicides. *Critical Reviews in Microbiology*, *42*(2), 245–261. Available from <https://doi.org/10.3109/1040841X.2014.929564>.
- Singh, B. K., Kuhad, R. C., Singh, A., Lal, R., & Tripathi, K. K. (1999). Biochemical and molecular basis of pesticide degradation by microorganisms. *Critical Reviews in Biotechnology*, *19*(3), 197–225. Available from <https://doi.org/10.1080/0738-859991229242>.
- Singh, S., Kumar, V., Chauhan, A., Datta, S., Wani, A. B., Singh, N., & Singh, J. (2018). Toxicity, degradation and analysis of the herbicide atrazine. *Environmental Chemistry Letters*, *16*(1), 211–237. Available from <https://doi.org/10.1007/s10311-017-0665-8>.
- Sohn, S. B., Kim, T. Y., Park, J. M., & Lee, S. Y. (2010). In silico genome-scale metabolic analysis of *Pseudomonas putida* KT2440 for polyhydroxyalkanoate synthesis, degradation of aromatics and anaerobic survival. *Biotechnology Journal*, *5*(7), 739–750. Available from <https://doi.org/10.1002/biot.201000124>.
- Sørensen, S. R., Simonsen, A., & Aamand, J. (2009). Constitutive mineralization of low concentrations of the herbicide linuron by a *Variovorax* sp. strain. *FEMS Microbiology Letters*, *292*(2), 291–296. Available from <https://doi.org/10.1111/j.1574-6968.2009.01501.x>.
- Souza, F. L., Saéz, C., Llanos, J., Lanza, M. R. V., Cañizares, P., & Rodrigo, M. A. (2016). Solar-powered electrokinetic remediation for the treatment of soil polluted with the herbicide 2,4-D. *Electrochimica Acta*, *190*, 371–377. Available from <https://doi.org/10.1016/j.electacta.2015.12.134>.
- Spina, F., Cecchi, G., Landinez-Torres, A., Pecoraro, L., Russo, F., Wu, B., . . . Persiani, A. M. (2018). Fungi as a toolbox for sustainable bioremediation of pesticides in soil and water. *Plant Biosystems*, *152*(3), 474–488. Available from <https://doi.org/10.1080/11263504.2018.1445130>.
- Strong, L. C., Rosendahl, C., Johnson, G., Sadowsky, M. J., & Wackett, L. P. (2002). *Arthrobacter aureescens* TC1 metabolizes diverse s-triazine ring compounds. *Applied and Environmental Microbiology*. Available from <https://doi.org/10.1128/AEM.68.12.5973-5980.2002>.
- Sun, J. Q., Huang, X., Chen, Q. L., Liang, B., Qiu, J. G., Ali, S. W., & Li, S. P. (2009). Isolation and characterization of three *Sphingobium* sp. strains capable of degrading isoproturon and cloning of the catechol 1,2-dioxygenase gene from these strains. *World Journal of Microbiology and Biotechnology*, *25*(2), 259–268. Available from <https://doi.org/10.1007/s11274-008-9888-y>.
- Sun, S., Sidhu, V., Rong, Y., & Zheng, Y. (2018). Pesticide pollution in agricultural soils and sustainable remediation methods: A review. *Current Pollution Reports*, *4*(3), 240–250. Available from <https://doi.org/10.1007/s40726-018-0092-x>.
- Suthers, P. F., Dasika, M. S., Kumar, V. S., Denisov, G., Glass, J. I., & Maranas, C. D. (2009). Genome-scale metabolic reconstruction of mycoplasma genitalium, iPS189. *PLoS Computational Biology*, *5*(2). Available from <https://doi.org/10.1371/journal.pcbi.1000285>.
- Sviridov, A. V., Shushkova, T. V., Ermakova, I. T., Ivanova, E. V., Epiktetov, D. O., & Leontievsky, A. A. (2015). Microbial degradation of glyphosate herbicides (review). *Applied Biochemistry and Microbiology*, *51*(2), 188–195. Available from <https://doi.org/10.1134/S0003683815020209>.
- Tasca, A. L., & Fletcher, A. (2018). State of the art of the environmental behaviour and removal techniques of the endocrine disruptor 3,4-dichloroaniline. *Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering*. Available from <https://doi.org/10.1080/10934529.2017.1394701>.
- Terzer, M., Maynard, N. D., Covert, M. W., & Stelling, J. (2009). Genome-scale metabolic networks. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *1*(3), 285–297. Available from <https://doi.org/10.1002/wsbm.37>.
- Thiele, I., & Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. Available from <https://doi.org/10.1038/nprot.2009.203>.
- Torre, A., Metivier, A., Chu, F., Laurens, L. M. L., Beck, D. A. C., Pienkos, P. T., . . . Kalyuzhnaya, M. G. (2015). Genome-scale metabolic reconstructions and theoretical investigation of methane conversion in *Methylobacterium buryatense* strain 5G(B1). *Microbial Cell Factories*, *14*(1), 1–15. Available from <https://doi.org/10.1186/s12934-015-0377-3>.
- Trigo, A., Valencia, A., & Cases, I. (2009). Systemic approaches to biodegradation. *FEMS Microbiology Reviews*, *33*(1), 98–108. Available from <https://doi.org/10.1111/j.1574-6976.2008.00143.x>.
- Tsoka, S., Simon, D., & Ouzounis, C. A. (2004). Automated metabolic reconstruction for *Methanococcus jannaschii*. *Archaea (Vancouver, BC)*, *1*(4), 223–229. Available from <https://doi.org/10.1155/2004/324925>.
- Valdez-Vazquez, I., Castillo-Rubio, L. G., Pérez-Rangel, M., Sepúlveda-Gálvez, A., & Vargas, A. (2019). Enhanced hydrogen production from lignocellulosic substrates via bioaugmentation with *Clostridium* strains. *Industrial Crops and Products*, *137*, 105–111. Available from <https://doi.org/10.1016/j.indcrop.2019.05.023>.
- Vats, S. (2015). *Herbicides: History, classification and genetic manipulation of plants for herbicide resistance*. Springer. Available from https://doi.org/10.1007/978-3-319-09132-7_3.
- Vázquez-Núñez, E., Molina-Guerrero, C. E., Peña-Castro, J. M., Fernández-Luqueño, F., & de la Rosa-Álvarez, M. G. (2020). Use of nanotechnology for the bioremediation of contaminants: A review. *Processes*. Available from <https://doi.org/10.3390/pr8070826>.
- Vongsangnak, W., Raethong, N., Mujchariyakul, W., Nguyen, N. N., Leong, H. W., & Laoteng, K. (2017). Genome-scale metabolic network of *Cordyceps militaris* useful for comparative analysis of entomopathogenic fungi. *Gene*, *626*, 132–139. Available from <https://doi.org/10.1016/j.gene.2017.05.027>.

- Wang, Q., & Xie, S. (2012). Isolation and characterization of a high-efficiency soil atrazine-degrading *Arthrobacter* sp. strain. *International Biodeterioration and Biodegradation*, 71, 61–66. Available from <https://doi.org/10.1016/j.ibiod.2012.04.005>.
- Wang, Y., Lai, A., Latino, D., Fenner, K., & Helbling, D. E. (2018). Evaluating the environmental parameters that determine aerobic biodegradation half-lives of pesticides in soil with a multivariable approach. *Chemosphere*, 209, 430–438. Available from <https://doi.org/10.1016/j.chemosphere.2018.06.077>.
- Wang, Y., Li, H., Feng, G., Du, L., & Zeng, D. (2017). Biodegradation of diuron by an endophytic fungus *Neurospora intermedia* DP8-1 isolated from sugarcane and its potential for remediating diuron-contaminated soils. *PLoS One*, 12(8), 1–18. Available from <https://doi.org/10.1371/journal.pone.0182556>.
- Wei, S., Jian, X., Chen, J., Zhang, C., & Hua, Q. (2017). Reconstruction of genome-scale metabolic model of *Yarrowia lipolytica* and its application in overproduction of triacylglycerol. *Bioresources and Bioprocessing*. Available from <https://doi.org/10.1186/s40643-017-0180-6>.
- Wu, X., Wang, W., Liu, J., Pan, D., Tu, X., Lv, P., ... Hua, R. (2017). Rapid biodegradation of the herbicide 2,4-dichlorophenoxyacetic acid by *Cupriavidus gilardii* T-1. *Journal of Agricultural and Food Chemistry*, 65(18), 3711–3720. Available from <https://doi.org/10.1021/acs.jafc.7b00544>.
- Xu, L., & Geelen, D. (2018). Developing biostimulants from agro-food and industrial by-products. *Frontiers in Plant Science*, 871, 1–13. Available from <https://doi.org/10.3389/fpls.2018.01567>.
- Xu, X. H., Liu, X. M., Zhang, L., Mu, Y., Zhu, X. Y., Fang, J. Y., ... Jiang, J. D. (2018). Bioaugmentation of chlorothalonil-contaminated soil with hydrolytically or reductively dehalogenating strain and its effect on soil microbial community. *Journal of Hazardous Materials*, 351, 240–249. Available from <https://doi.org/10.1016/j.jhazmat.2018.03.002>.
- Xu, X., Zarecki, R., Medina, S., Ofaim, S., Liu, X., Chen, C., ... Freilich, S. (2019). Modeling microbial communities from atrazine contaminated soils promotes the development of biostimulation solutions. *ISME Journal*, 13(2), 494–508. Available from <https://doi.org/10.1038/s41396-018-0288-5>.
- Yang, F., Jiang, Q., Zhu, M., Zhao, L., & Zhang, Y. (2017). Effects of biochars and MWNTs on biodegradation behavior of atrazine by *Acinetobacter lwoffii* DNS32. *Science of the Total Environment*, 577, 54–60. Available from <https://doi.org/10.1016/j.scitotenv.2016.10.053>.
- Yang, X., Wei, H., Zhu, C., & Geng, B. (2018). Biodegradation of atrazine by the novel *Citricoccus* sp. strain TT3. *Ecotoxicology and Environmental Safety*, 147, 144–150. Available from <https://doi.org/10.1016/j.ecoenv.2017.08.046>.
- Yang, Y., Pratap Singh, R., Song, D., Chen, Q., Zheng, X., Zhang, C., ... Li, Y. (2020). Synergistic effect of *Pseudomonas putida* II-2 and *Achromobacter* sp. QC36 for the effective biodegradation of the herbicide quinclorac. *Ecotoxicology and Environmental Safety*, 188. Available from <https://doi.org/10.1016/j.ecoenv.2019.109826>.
- Yao, X. F., Khan, F., Pandey, R., Pandey, J., Mourant, R. G., Jain, R. K., ... Pandey, G. (2011). Degradation of dichloroaniline isomers by a newly isolated strain, *Bacillus megaterium* IMT21. *Microbiology (Reading, England)*, 157(3), 721–725. Available from <https://doi.org/10.1099/mic.0.045393-0>.
- Ye, C., Xu, N., Chen, H., Chen, Y. Q., Chen, W., & Liu, L. (2015). Reconstruction and analysis of a genome-scale metabolic model of the oleaginous fungus *Mortierella alpina*. *BMC Systems Biology*, 9(1), 1–11. Available from <https://doi.org/10.1186/s12918-014-0137-8>.
- You, I. S., & Bartha, R. (1982). Stimulation of 3,4-dichloroaniline mineralization by aniline. *Applied and Environmental Microbiology*, 44, 678–681.
- Yu, X. M., Yu, T., Yin, G. H., Dong, Q. L., An, M., Wang, H. R., & Ai, C. X. (2015). Glyphosate biodegradation and potential soil bioremediation by *Bacillus subtilis* strain Bs-15. *Genetics and Molecular Research*, 14(4), 14717–14730. Available from <https://doi.org/10.4238/2015.November.18.37>.
- Zanardini, E., Arnoldi, A., Boschini, G., D'Agostina, A., Negri, M., & Sorlini, C. (2002). Degradation pathways of chlorsulfuron and metsulfuron-methyl by a *Pseudomonas fluorescens* strain. *Annals of Microbiology*, 52(1), 25–37.
- Zeng, S., Qin, X., & Xia, L. (2017). Degradation of the herbicide isoproturon by laccase-mediator systems. *Biochemical Engineering Journal*, 119, 92–100. Available from <https://doi.org/10.1016/j.bej.2016.12.016>.
- Zhang, J., Zheng, J. W., Liang, B., Wang, C. H., Cai, S., Ni, Y. Y., ... Li, S. P. (2011). Biodegradation of chloroacetamide herbicides by *Paracoccus* sp. FLY-8 in vitro. *Journal of Agricultural and Food Chemistry*. Available from <https://doi.org/10.1021/jf104695g>.
- Zhang, L., Hang, P., Hu, Q., Chen, X. L., Zhou, X. Y., Chen, K., & Jiang, J. D. (2018). Degradation of phenylurea herbicides by a novel bacterial consortium containing synergistically catabolic species and functionally complementary hydrolases. *Journal of Agricultural and Food Chemistry*, 66(47), 12479–12489. Available from <https://doi.org/10.1021/acs.jafc.8b03703>.
- Zhang, X., Gao, Y., Zang, P., Zhao, Y., He, Z., Zhu, H., ... Zhang, L. (2019). Study on the simultaneous degradation of five pesticides by *Paenibacillus polymyxa* from *Panax ginseng* and the characteristics of their products. *Ecotoxicology and Environmental Safety*, 168, 415–422. Available from <https://doi.org/10.1016/j.ecoenv.2018.10.093>.
- Zhang, Y., Chen, H., Liu, J., Geng, G., Liu, D., Geng, H., & Xiong, L. (2018). Genome sequencing and biodegradation characteristics of the n-butyl benzyl phthalate degrading bacterium *Rhodococcus* sp. HS-D2. *International Biodeterioration and Biodegradation*, 128, 56–62. Available from <https://doi.org/10.1016/j.ibiod.2016.08.024>.
- Zhelezova, A., Cederlund, H., & Stenström, J. (2017). Effect of biochar amendment and ageing on adsorption and degradation of two herbicides. *Water, Air, and Soil Pollution*, 228(6). Available from <https://doi.org/10.1007/s11270-017-3392-7>.
- Zhu, J., Fu, L., Jin, C., Meng, Z., & Yang, N. (2019). Study on the isolation of two atrazine-degrading bacteria and the development of a microbial agent. *Microorganisms*, 7(3), 1–11. Available from <https://doi.org/10.3390/microorganisms7030080>.
- Zhu, Y., Czauderna, T., Zhao, J., Klapperstueck, M., Maifiah, M. H. M., Han, M. L., ... Li, J. (2018). Genome-scale metabolic modeling of responses to polymyxins in *Pseudomonas aeruginosa*. *GigaScience*. Available from <https://doi.org/10.1093/gigascience/giy021>.
- Zuanazzi, N. R., Ghisi, N. de C., & Oliveira, E. C. (2020). Analysis of global trends and gaps for studies about 2,4-D herbicide toxicity: A scientometric review. *Chemosphere*, 241. Available from <https://doi.org/10.1016/j.chemosphere.2019.125016>.

This page intentionally left blank

Chloroplast genome and plant–virus interaction

Parampreet Kaur¹, Tanvi Kaila², Manmohan Dhkal¹ and Kishor Gaikwad²

¹*School of Organic Farming, Punjab Agricultural University, Ludhiana, India,* ²*ICAR-National Institute for Plant Biotechnology, New Delhi, India*

26.1 Introduction

The origin of first photosynthetic eukaryote dates back to more than 1000 million years ago. The prime attainment of a cyanobacterial endosymbiont by a eukaryotic host, eventually catalyzed formation of chloroplast containing green algae and subsequently higher plants. Chloroplasts are often known as metabolic centers of plants, as they are the chief organelles involved in photosynthesis and biosynthesis of metabolites such as, vitamins, phytohormones, amino acids, and nucleotides. They are involved in assimilation of nitrogen and sulfur and are also associated with the synthesis of various metabolites which are involved in defense against pathogens and abiotic stresses.

Among different plant pathogens, plant viruses are one of the most widely spread and economically important pathogens. Basically they are nucleoprotein containing obligate parasites that multiply inside the living host cell by using their components and resulting in the development of symptoms on infected plants that ultimately causes diseases. Nearly all the crop plants, grown either for food or fiber, get affected by one or more than one plant viruses. Though the cultivated crops are common hosts for most of the viruses, several reports of virus infection on wild species are also available, which acts as both host and reservoir for different viruses (Hull, 2014). The recognition of viruses as an infectious entity was reported by M.W. Beijerinck in late 1890s, very early records of virus infection in crop plant are also known. A Japanese poem written by Empress Koken in CE 752 was found to be the earliest known record of virus disease where infection of virus on tulip flower leads to the development of color break type of symptoms. Further, the work of Albert Mayer in 1886 established the infectious nature of “*MOSAIKKRANKHEIT*” [later known to be caused by *Tobacco mosaic virus* (TMV)], a disease of tobacco which can be transmitted from a diseased to healthy plant by inoculation with leaf extracts from infected plants. In 1892, Dmitri Ivanovsky proved that the aforementioned disease of tobacco was not a bacterial infection as the sap of infected plant remained infectious after passing it through bacterial proof filter paper. In 1898, M.W. Beijerinck gave the term “*Contagium Vivum Fluidum*” to the agent present in the infectious fluid and responsible for the disease in tobacco plant (Hull, 2014). Further discoveries of viruses as plant pathogens is attributed to the repertoire of classical and advanced research work done around the globe in different aspects of virology, that is, detection and diagnosis of plant pathogenic viruses, classification, symptomatology, mechanism of infection, gene cloning, etc. Application of high throughput Next Generation Sequencing (NGS) technologies has further aided the progress. Ever since the decoding of first viral genome using omics, discovery, and identification of novel viruses/viroids, their characterization and diagnostic methodology has been refined tremendously, thus making a mammoth impact in plant virology (Barba, Czosnek, & Hadidi, 2014; Blawid, Silva, & Nagata, 2017; Hadidi, Flores, Candresse, & Barba, 2016; Pecman et al., 2017).

Plant viruses generally encode very few proteins, owing to their small genome sizes and thus largely bank on the host cellular machinery for their propagation and spread. Effects of viral infection include necrosis, stunning, and plant chlorosis. Leaf chlorosis is the most common viral symptom which is generally associated with reduced photosynthetic activity. Past studies have reported that changes in expression of chloroplast related genes and chloroplast components and structure due to viral infection, were responsible for development of viral symptoms in plants (Li, Cui, Cui, & Wang, 2016; Manfre, Glenn, Nunez, Moreau, & Dardick, 2011; Revers & García, 2015; Xu & Nagy, 2010). Chloroplast and its factors are known to interact with or become targets of viruses and thereby favor their replication,

movement and symptom development. It is known that chloroplast is a target of choice for viruses and it undergoes colossal damage both functionally and structurally. Other modifications coupled with leaf chlorosis such as reduced chlorophyll pigmentation (Balachandran, Osmond, & Daley, 1994; Wang et al., 2018), decreased expression of *CPRGs*, that is, Chloroplast and Photosynthesis-Related Genes encoded by nuclear genome (Dardick, 2007; Das, Lin, & Wong, 2018; Mochizuki, Ogata, Hirata, & Ohki, 2014), alterations in chloroplast functioning and aberrant structures (Bhat et al., 2012; Otulak, Chouda, Bujarski, & Garbaczewska, 2015), and accumulation of nitric oxide (Mwaba & Rey, 2017), implies indispensable interactions between virus and the chloroplast (Zhao, Zhang, Hong, & Liu, 2016). For instance, a 2b mutant strain (pepo strain) of Cucumber mosaic virus (CMV) possesses point mutations in its coat protein and represses the expression of *CPRGs* in the host plant (Mochizuki et al., 2014). In fact, there are reports which provides evidence for genetic material exchange among the plants and viruses and hence evolution of host and pathogen relationship between the two. For example in the tobacco nuclear genome, copies of the geminiviral replication protein as well as the chromosomal material of partitiviruses and totiviruses have been found to be incorporated (Liu et al., 2010). Another example is of CMV Y-Sat virus, which infects *Nicotiana* species with the help of small interfering (siRNA)-directed RNA and silences the host chlorophyll biosynthetic gene (*CHLI*). The aforementioned mechanism is known to be aided by the presence of a sequence (22-nucleotide) in “yellow region” of CMV Y-Sat complementary to a sequence in *CHLI*. Further, *Nicotiana* species lacking this complementary stretch exhibit considerable resistance to the viral infection (Shimura et al., 2011; Smith, Eamens, & Wang, 2011).

26.2 Chloroplast genome

26.2.1 Structure and gene content

Plant chloroplasts generally consists of a typical structure, that is, a quadripartite structure composed of 1 large single copy (LSC), 1 small single copy (SSC), and 2 inverted repeats (IRs) region. Size of these regions may vary among different plant species (Fig. 26.1). For instance, the IR region spans 12–75 kb of the region separating LSC and SSC region which varies between 80–90 kb and 16–27 kb of size respectively. Generally, among the diverse species the chloroplast genome is known to be conserved, although variation in the length of intergenic spacers and the events like contraction, expansion and loss of IR regions results in the observed variations in the size of the chloroplast genome among different species. For instance, chloroplast genome size ranges from 107 kb in *Cathaya argyrophylla* to 218 kb in *Pelargonium* (Daniell, Choun, Ming, & Wan, 2016). Although, chloroplast genome comprises of a single circular molecule, there are also studies reporting the existence of linear form of chloroplast genomes and it has been observed that the percentage of each form varies within the cells, among different reports (Oldenburg & Bendich, 2015, 2016). The chloroplast genome contains 110–130 genes, which comprises of 80–90 protein coding genes, 30–31 tRNAs and 4 rRNAs. Majority of the chloroplast proteins are nuclear encoded and are transported to the chloroplast with the help of transit peptide, a short amino acid sequences present on the N-terminal of protein (Jarvis & Soll, 2001; Leister, 2003).

Although the chloroplast genome is conserved, differences in gene copy number and synteny has also been reported. In the course of evolution, various genes from the plastid have got transferred to the nuclear genome. One such example is of *infA* gene (chloroplast encoded), there are reports showing the translation of *infA* gene in the cytosol of some plants (*Arabidopsis thaliana* and *Glycine max*) and subsequent transportation of the protein to the chloroplast with the help of transit sequence. Similarly, genes like *rpl22* and *rpl32* have been moved to the nuclear genome over the course of evolution. Considering *NDH* family, partial to complete loss of this gene family has been reported. For example, pine and orchids lack function *ndh* genes in their plastome (Lin et al., 2015). Other genes like *accD*, *rps16*, *rpl23*, *rpl33*, *psaI*, and *ycf4* have also been reported to be lost from the plastid genome (Jansen et al., 2007; Magee et al., 2010). Genes like *ndhF* and *ycf2* have been observed to be lost repeatedly in the course of evolution from various angiosperms (Sato, Nakamura, Kaneko, Asamizu, & Tabata, 1999; Shinozaki et al., 1986).

On the other hand, duplication and pseudonization of plastid genes has also been reported. For instance, in maize and rice, *ycf2* pseudogene present in the chloroplast genome governs cell viability; *ycf2* is reported to be present in plastome of various land plants (Hiratsuka et al., 1989; Maier, Neckermann, Igloi, & Kössel, 1995). Similarly, *rpl23* is a pseudogene present in spinach plastid genome and *infA* is present as a pseudogene in tobacco and *Oenothera elata* chloroplast genome (Thomas, Massenet, Dorne, & Briat, 1988). Likewise, some tRNA genes, *ycf2*, *psbA*, and *rpl23* have been reported to undergo duplication in some plastomes.

Among the angiosperms, extensive rearrangements in the chloroplast genomes have been reported to those belonging to the fabaceae family compared to others (Cai et al., 2008; Guo et al., 2007; Jansen, Wojciechowski, Sanniyasi,

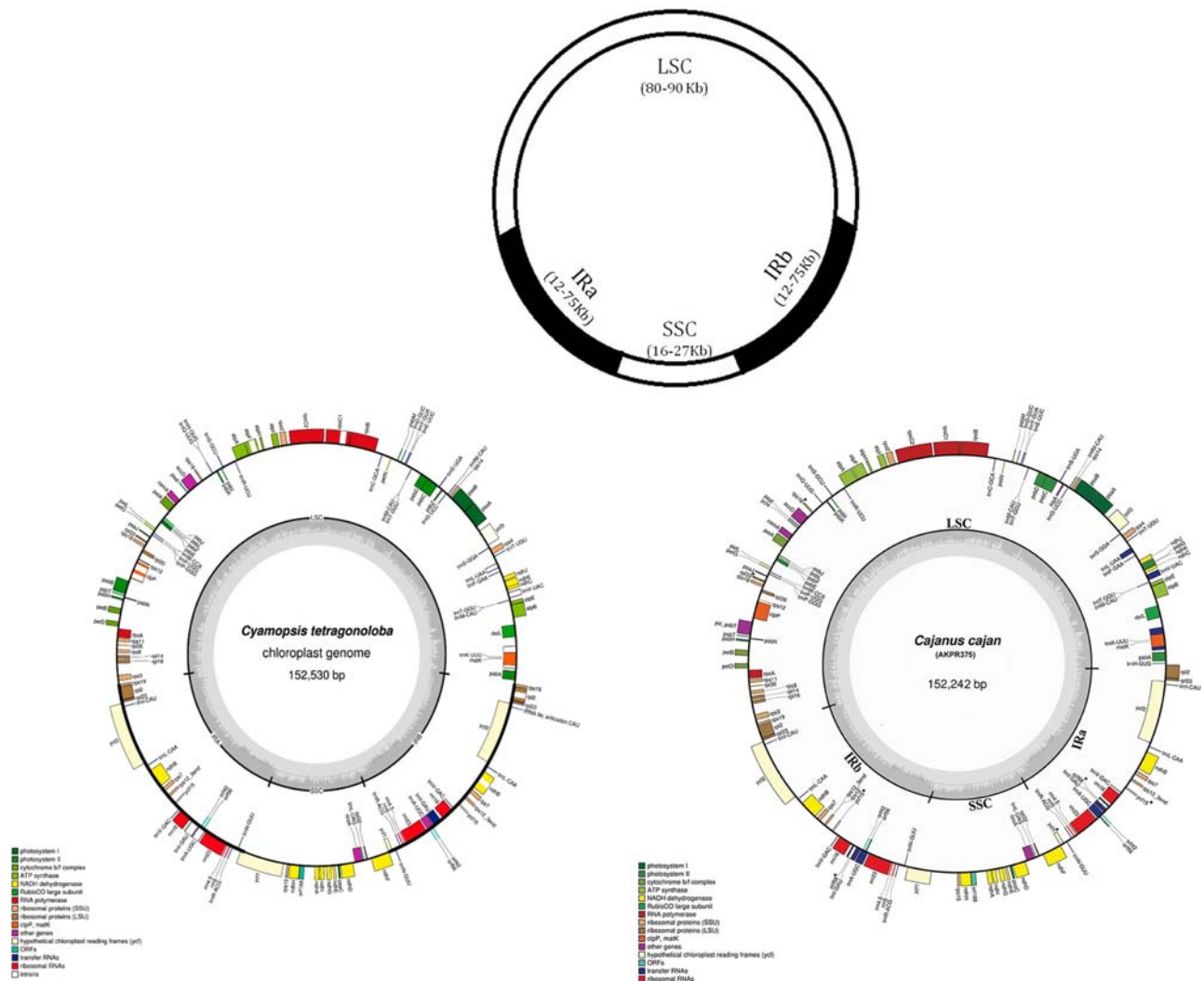


FIGURE 26.1 General structure of legume chloroplast genome (top center), chloroplast genome of *Cyamopsis tetragonoloba* (Kaila et al., 2017) and *Cajanus cajan* (Kaila et al., 2016).

Lee, & Daniell, 2008; Kato, Kaneko, Sato, Nakamura, & Tabata, 2000). Moreover, in large clades of legumes a complete loss of IRs has happened and tribes like *Trifolieae*, *Fabeae*, *Cicereae*, *Galegeae*, *Carmichaelieae*, and *Hedysareae* belonging to this clade are known as Inverted Repeat Lacking Clade (IRLC) (Wojciechowski, Sanderson, Steele, & Liston, 2000). Various rearrangements reported in legume genomes like 50 kb inversion in the LSC region (Palmer & Thompson, 1982; Palmer, Osorio, & Thompson, 1988), 78 kb rearrangement reported in the LSC region of *Phaseolus* and *Vigna* (Bruneau, Doyle, & Palmer, 1990; Guo et al., 2007; Tangphatsornruang et al., 2010), and a 36 kb inversion within 50 kb inversion, newly reported in lupines and other genisotoids (Martin et al., 2014), are believed to be the result of unstable chloroplast genome. It has been observed that the loss of IR has made the chloroplast genome more liable to rearrangements (Doyle, Doyle, & Palmer, 1995; Palmer & Thompson, 1982).

Various instances of intron loss have also been reported in angiosperms. Protein coding genes from various species, that is, chickpea (Jansen et al., 2008), cassava (Daniell et al., 2008) and barley (Saski et al., 2007) have lost their introns in the course of evolution. Intron loss has also been observed in *atpF* gene and recombination between an intron lacking and intron bearing copy of cDNA has been explained as the mechanism behind intron loss in Malpighiales. Other genes namely, *rpl2*, *rps12*, and *rps16* (ribosomal proteins), *rpoC2*, and *clpP* have undergone intron losses (Jansen et al., 2007). Loss of intron in *clpP* gene also marks for the monophyly of IRLC (*Trifolium*, *Pisum sativum*, *Lathyrus sativus*, *Cicer* and *Medicago*) (Jansen et al., 2008).

26.2.2 Genomic advances

With the advent of high-throughput sequencing technology, there has been rapid advancement in the discipline of chloroplast genomics, genetics and engineering. The first plastome genomes to be sequenced were those of tobacco (Shinozaki et al., 1986) and liverwort (Ohyama et al., 1986) and since then 4645 chloroplast genomes have been sequenced and are available at NCBI organelle genome resources (<https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/>).

With the availability of complete chloroplast genome sequences, deep insights into the evolutionary relationships of plants among the phylogenetic clades have been established. The analysis of different plastome sequences has also revealed significant sequence and structural variations among them. The instances of gene transfer from the chloroplast genome, to the nuclear and mitochondrial genomes have been revealed with the help of plastome sequence analysis and have helped in deciphering the relationship among the three plant genomes (chloroplast, mitochondrial, and nuclear). From providing protection against various abiotic and biotic stresses to progress in vaccine development, chloroplast genomics have proved to have important applications as well. As compared to earlier methods such as, rolling circle amplification (Jansen et al., 2008; Lee et al., 2006; Ruhlman et al., 2006; Saski et al., 2007) or cloning into Bacterial Artificial Chromosome (BAC) vectors or Fosmids (Jansen, Saski, Lee, Hansen, & Daniell, 2011; Saski et al., 2005; Wolfe, Morden, & Palmer, 1992), NGS has provided faster and cheaper means to sequence the chloroplast genome. Moore and colleagues (2006) were the first to explore the field of NGS (454 GS-20 systems) to decode the chloroplast genome sequences of *Nandina domestica* and *Platanus occidentalis*. Nowadays, the major NGS platform used for chloroplast genome sequencing is Illumina. As sequencing with Illumina produces short reads, it can be combined with PacBio platform (third generation sequencer) which produces long reads.

26.2.3 Bioinformatic approaches and plastomes

Chloroplast genome assembly done by using both short and long reads results in high accuracy. In recent years, a lot of efforts have been put in for the amelioration of chloroplast genomics. A lot of bioinformatics tools have been developed starting from assembly to visualization (Fig. 26.2). If we start from the organelle genome assembly, then tools like ORGanelle ASseMbler, NOVOPlasty, Canu, CLC workbench are available. ORGanelle ASseMbler (<http://python-hosted.org/ORG.asm/>) is a command line software tool, which uses short reads for organelle genome assembly. NOVOPlasty (Dierckxsens, Mardulyn, & Smits, 2016) is a helpful tool for de novo assembly of chloroplast genomes. Canu (Koren et al., 2017) is a tool which is useful in assembling reads generated from PacBio or Oxford Nanopore platform.

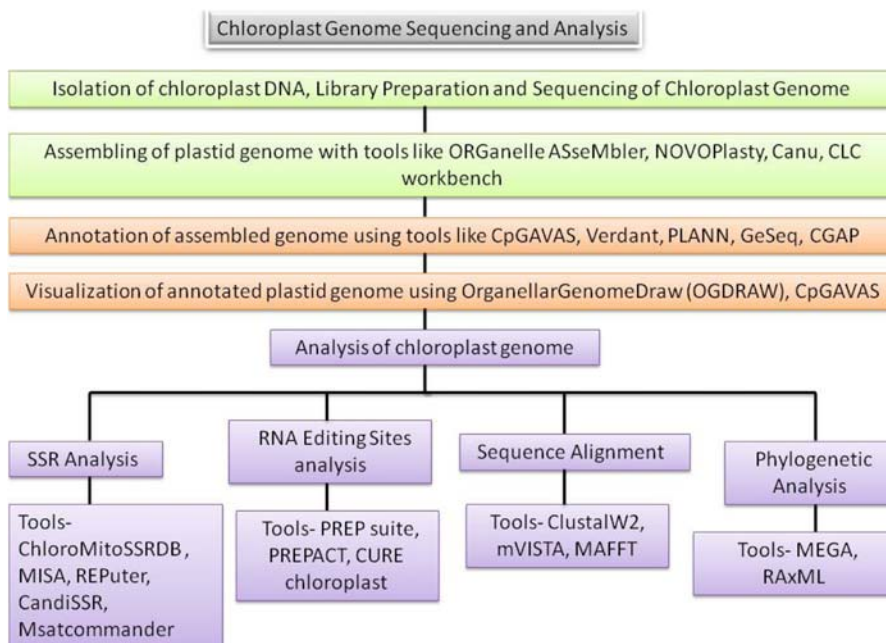


FIGURE 26.2 General approaches to chloroplast genome analysis using various bioinformatics tools.

The CLC Genomics workbench (<https://www.qiagenbioinformatics.com/>) is commercial software that assists in de novo as well as reference based assembly. The next step, the functional annotation of the chloroplast genome is a key stage in the whole process. Earlier DOGMA (Wyman, Jansen, & Boore, 2004) was the most frequently used tool for this purpose but it is no longer used. Other tools namely, CpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis and GenBank Submission) (Liu et al., 2012), Verdant, PLANN (Plastome Annotator) (Huang & Cronk, 2015), GeSeq (Tillich et al., 2017) and CGAP (Chloroplast Genome Analysis Platform) (Cheng, Zeng, Ren, & Liu, 2013) are used for organelle genome annotation. Display of genes coding proteins, tRNAs and rRNAs, and boundaries defining different regions of the plastome is a pivotal feature among different steps of chloroplast genome analysis. The Organellar Genome Draw (OGDRAW) is a tool which allows the users to generate good quality graphs (circular and linear) (Lohse, Drechsel, & Bock, 2007; Lohse, Drechsel, Kahlau, & Bock, 2013).

Several genomic resources to mediate plant–virus interactions such as Phytozome (<https://phytozome.jgi.doe.gov/pz/portal.html>), viruSITE (<http://www.virusite.org/index.php>), Descriptions of Plant Viruses (DPV, <http://www.dpvweb.net/seqs/plantviruses.php>) Virus-host DB (<https://www.genome.jp/virushostdb/index/virus/all>), PlatGDB (<http://www.platgdb.org/>), PLAZA (<https://bioinformatics.psb.ugent.be/plaza/>), etc., are freely available. One can also explore different aspects like SSR (Simple Sequence Repeats) mining and posttranscriptional modifications using genome sequencing in chloroplasts. The chloroplast genome houses conserved gene regions that have led to the development of molecular markers and their analysis across different species. The MISA perl script (Thiel, 2003) allows the extraction of SSRs present in the plastome. ChloroMitoSSRDB (Sablok et al., 2013) is a repository which assists in large scale visualization of SSRs across the plastomes. Other tools like, PREP suite (Predictive RNA Editor for Plants) (Mower, 2009) and PREPACT (Plant RNA Editing—Prediction & Analysis Computer Tool) (Lenz & Knoop, 2013) allows to explore the aspect of RNA Editing which entails cytidine to uridine conversion and constitutes an important aspect of RNA maturation.

26.2.4 Status of chloroplast genome sequencing in plants

With the increase in availability of resources, there has been substantial increase in the number of sequenced genomes being deposited in NCBI. Till date, 4645 chloroplast genome sequences have been submitted to NCBI organelle genome resources. Organelle genomes play an important role in DNA barcoding studies (Dong, Liu, Yu, Wang, & Zhou, 2012), phylogenetic studies (Provan, Powell, & Hollingsworth, 2001) and species identification (Li et al., 2015). One can explore various aspects like genome organization, gene number, comparison of gene order among different species and RNA editing status. Also, SSR markers can be developed from the plastome and the same can be studied for cross-transferability across the species (Saxena et al., 2019). There has been a tremendous increase in number of chloroplast genomes being submitted in NCBI for past few years. Taking a look at some families of land plants, then Poaceae (573) contains maximum number of species whose chloroplast genomes has been sequenced and submitted to NCBI, followed by Fabaceae (155), Malvaceae (57), and Piperaceae (5) families. Few important crop plants whose chloroplasts genomes have been sequenced includes *Oryza sativa* (Hiratsuka et al., 1989), *Raphanus sativus* L. Jeong, Chung, Mun, Kim, and Yu (2014), *Vigna radiata* (Tangphatsornruang et al., 2010), *Glycine max* (Saski et al., 2005), *Cajanus cajan* (Kaila et al., 2016), *Lotus japonicus* (Kato et al., 2000) and *Cyamopsis tetragonoloba* L. Kaila et al. (2017) (Table 26.1). The increased availability of chloroplast genomes has opened the doors for another field, such as, Chloroplast Genetic Engineering. Transformation of chloroplast genomes has offered various advantages over nuclear genomes. High levels of foreign protein accumulation as polyploidy nature of chloroplast genomes permits the introduction of thousands of copies of foreign gene (De Cosa, Moar, Lee, Miller, & Daniell, 2001), lack of gene silencing in genetically engineered chloroplasts, containment of transgene by maternal inheritance (Hagemann, 2004; Zhang & Liu, 2003), and absence of position and pleiotropic effects (Jin & Daniell, 2015) are some of the advantages.

With the help of chloroplast engineering numerous possibilities have arisen and can lead to the development of crops exhibiting resistance to insects, bacterial, viral, and fungal diseases. Production of biopharmaceuticals, industrial enzymes, bio fuels, and various antigens have been facilitated with the help of plastid engineering (Bock, 2007; Bock & Warzecha, 2010; Clarke & Daniell, 2011; Daniell, Singh, Mason, & Streatfield, 2009; Daniell et al., 2016).

26.3 Viral infection symptoms in plants

Directly or indirectly plant pathogenic virus affects majority of cell organelles during their replication and movement that causes various histological changes that is, necrosis, hypoplasia and hyperplasia. These histological changes either singly or in combination produce various macroscopic symptoms which includes yellowing, puckering, blistering, leaf

TABLE 26.1 List of few sequenced chloroplast genomes.

S. No.	Species	Chloroplast genome size	Sequencing platform	References
1	<i>Paphiopedilum delenatii</i>	160,955 bp	Illumina HiSeq	Vu et al. (2020)
2	<i>Metasequoia Glyptostroboides</i>	131,887	Illumina Miseq	Chen et al. (2015)
3	Globe artichoke	152,529	Illumina GAllx	Curci, De Paola, Danzi, Vendramin, and Sonnante (2015)
4	<i>Acacia ligulata</i>	158,724	Illumina Hiseq 2000	Williams, Boykin, Howell, Nevill, and Small (2015)
5	<i>Lupinus luteus</i>	151,894	Illumina HiSeq 2000	Keller et al. (2017)
6	<i>Ipomoea batata</i>	161,303	Illumina Hiseq 2000	Yan et al. (2015)
7	<i>Ananas comosus</i>	159,636	Illumina and PacBio RSII	Nashima et al. (2015)
8	<i>Fragaria _ananassa</i> "Benihoppe"	155,549	Illumina HiSeq 2500	Cheng et al. (2017)
9	<i>Capsicum pubescens</i>	157,390	Illumina Hiseq 2500	D'Agostino et al. (2018)
10	<i>Panax Quinquifolius</i>	156,359	Roche 454 GS FLX and Illumina short-read	Han, Li, Liu, and Gao (2016)
11	<i>Pinus taeda</i> L.	121,531	Illumina Hiseq 2000	Asaf, Khan, Shahzad, Lubna, and Kang (2018)
12	<i>Fagus crenata</i>	158,227	Illumina Hiseq 2000	Worth, Liu, Wei, and Tomaru (2019)
13	<i>Raphanus sativus</i> L.	153,368	Illumina HiSeq1000, Roche/454 GS-FLX 129 Plus, and PacBio RS II	De Cosa et al. (2001)
14	<i>Cajanus cajan</i> (L.) Millspaugh	152,201	Roche 454 GS FLX	Leister (2003)
15	<i>Cajanus scarabaeoides</i> (L.) Thouars	152,242	Roche 454 GS FLX	Leister (2003)
16	<i>Cyamopsis tetragonoloba</i> L.	152,530	Illumina Hiseq 1000	Kaila et al. (2017)

curl, interveinal chlorosis, leaf deformation, mottling, stunted plant growth, and necrosis. (Fig. 26.3). Further, among different organelles, chloroplasts are most frequently infected during virus infection (Balachandran et al., 1994; Herbers et al., 2000; Kyselakova et al., 2011; Mandahar & Garg, 1972; Reiner & Beachy, 1989), which results in symptoms such as abnormal plant size, mosaic, yellows/chlorosis, and necrosis. Various viral infection symptoms include the following:

- 1. Abnormal plant size:** Dwarfing and stunting of infected plant is one of the most common symptoms produced as result of virus infection. Intensity of this symptom is commonly related to the degree of other symptoms, especially in which leaf chlorophyll losses are encountered. Stunting symptoms results due to the reduction in internode length and leaf size of the plants.
- 2. Mosaic:** Development of dark green and light green area pattern on leaf surface is called as mosaic and it is also a common symptom observed in virus infected plants. Almasi, Harsanyi, and Gaborjanyi (2001) observed that virus infection in the plant causes various degree of chloroplast damage in clustered mesophyll cells that leads to the production of mosaic like symptoms.

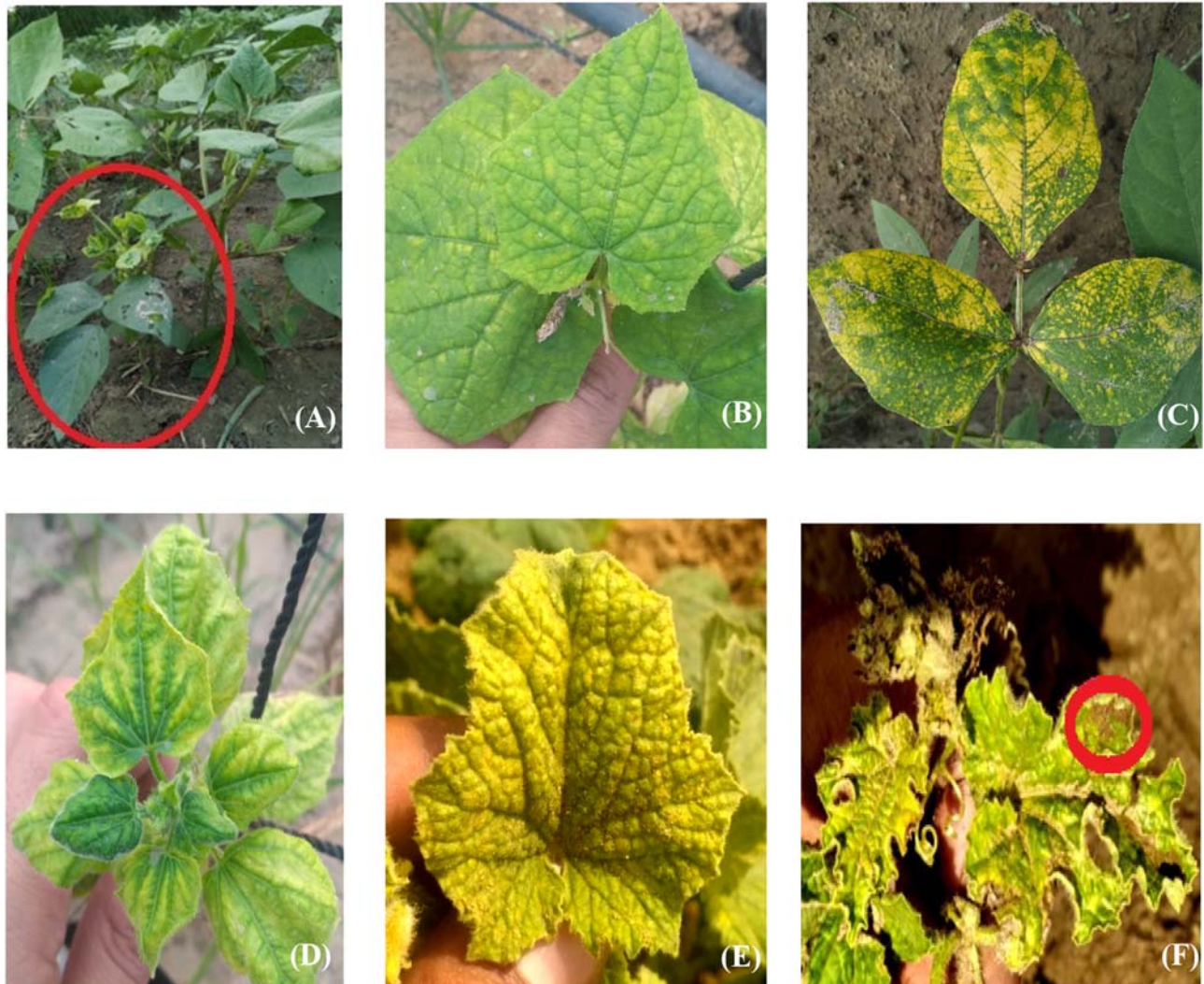


FIGURE 26.3 Major viral infection symptoms related to chloroplast infection (A) Stunting of moong bean plant due to *Moong bean yellow mosaic virus* infection (B) Mosaic symptom on cucumber plant (C) Mosaic on moong bean due to *Moong bean yellow mosaic virus* infection (D) Chlorosis symptom on cucumber due to the infection of *Tomato leaf curl New Delhi virus* (E) Yellows symptom on Muskmelon due to the infection of *Tomato leaf curl New Delhi virus* (F) Necrosis symptoms produced on tinda due to infection of *Groundnut bud necrosis virus* (Dhkal M personal photographs).

3. **Yellows/Chlorosis:** Yellows/chlorosis is one of the most predominant symptom of viral infection manifested in terms of chloroplast structural changes and altered pigmentation; depleted photosynthetic activities, etc. Vein clearing/yellowing in young leaves are the first sign of viral infection in yellows type symptoms that are then followed by general yellowing and reddening of the infected leaves. Several workers reported development of chlorosis symptoms following the virus infection (Manfre et al., 2011; Rahoutei, Garcia-Luque, & Baron, 2000; Roberts & Wood, 1982). Choudhury et al. (2019) observed that infection of *Barley yellow dwarf virus* in wheat genotypes leads to significant reduction in leaf chlorophyll content that results in the development of strong yellows type symptoms. Thus photosynthetic inhibition via disrupting components of chloroplast appears to be one of the conserved lines of attack adopted by virus pathogens to establish an ideal niche for themselves and spreading the infection.
4. **Necrosis:** Generally plant viruses don't kill the host cell due to its obligate parasitic nature. But certain viruses, such as those belonging to *Tomato spotted wilt virus* group are reported to kill their host cells/tissue by necrosis. In case of necrosis, death of plant tissue and organ occur but in case of severe infection the whole plant can die. Further, necrosis also affects the chlorophyll content of the infected plant to some extent. For instance, reduced chlorophyll

content in peanut leaves with necrosis symptom as compared to asymptomatic leaves has been reported (Rowland, Dornera, Sorensena, Beasley, & Todd, 2005).

26.4 Role of chloroplasts in plant–virus life cycle

Considering the relatively small sizes of the virus genomes, capturing of the host plant cell components to facilitate viral replication and their movement is inevitable. Though there is a constant tug-of-war between defense responses mediated by host cells in terms of hypersensitive response (HR)/Post transcriptional gene silencing etc. and suppression of viral gene silencing; a successful viral infection overcomes various defense barriers of a plant cell and triggers disease symptoms. Participation of various chloroplast constituents at different stages of virus infection cycle is reported during several intensive research studies conducted across globe. For instance, Xiang, Kakani, Reade, Hui, and Rochon (2006) observed the association of chloroplast in viral uncoating which is a major step in virus replication. Chloroplast consists of various compartments and membrane contents that are suitable for plant virus replication and helpful for them in evading plant RNA-mediated defense response (Ahlquist, Noueir, Lee, Kushner, & Dye, 2003; Dreher, 2004; Torrance, Cowan, Gillespie, Ziegler, & Lacomme, 2006). Following virus infection, “virus factories” are established at specific sites inside the cell for their efficient replication and movement, which is must to cause disease on the host plant (Zhao et al., 2016). Virus factories are that recruit some specific organelles for their build up excluding majority of host organelles and proteins. Majorly mitochondria, cytoplasmic membranes and cytoskeleton components of the host plant cells are involved in the formation of virus factories, which plays a prominent role in virus replication (Asurmendi, Berg, Koo, & Beachy, 2004).

Viral replication complexes (VRCs) are a major site for the progeny virus production and are commonly associated with chloroplast envelopes (includes cytoplasmic invaginations and peripheral vesicles) in majority of viral infections to probably prevent the recognition of viral RNAs from RNA-Silencing machinery of host (De Graaff, Coscoy, & Jaspars, 1993; Dreher, 2004; Torrance et al., 2006). Replication of viruses and development of subsequent virion assembly take place in chloroplast as different viral factors viz. genomic RNAs and viral protein, mediates chloroplast targeting of VRCs (Jakubiec et al., 2004; Prodhomme, Jakubiec, Tournier, Drugeon, & Jupin, 2003; Torrance et al., 2006). In chloroplast, there are some special components that help in chloroplast targeting of VRCs. For instance, lipid component of chloroplast membrane are associated with Pomovirus PMTV-TGB₂ and helps the viral RNA to stay at chloroplast membrane for replication (Cowan et al., 2012). Certain chloroplast proteins namely Chloroplast protein phosphoglycerate kinase (ChlPGK) and heat shock proteins (Hsp90) play an important role in virus replication (Budziszewska & Steplowska, 2018). In replication of bamboo mosaic virus, these protein along with eEF1a, glutathione S-transferase NbGSTU4 (from *N. benthamiana*), and exonuclease XRN4 are known to interact with 3' end of the viral genomic RNA during early stages of replication (Chen et al., 2017; Huang, Chen, & Tsai, 2017; Mower, 2009).

Further, to ensure virus survival in a host cell, its movement from one cell to another are crucial in its systematic spread and developing infectious symptoms. To fulfill these, the virus modifies various components of cell, that is, plasmodesmata, endomembrane system and cytoskeleton. Plant endomembrane system consists of many interrelated membranes and various organelles like chloroplast, mitochondria, endoplasmic reticulum, vacuole, endosomes, peroxisomes, nuclear envelopes, plasmamembrane and vesicles (Chen et al., 2012; Morita & Shimada, 2014). In the cytoskeleton, microtubule and actin filaments are two main components that are involved in proper positioning of endomembrane system and movement of constituents through them (Toivola, Strnad, Habtezion, & Omary, 2010; Wang & Hussey, 2015). Movement of virus requires various proteins known as “Movement Proteins” (MP) for its transport through host plant symplastic routes (Lazarowitz & Beachy, 1999; Wolf, Lucas, Deom, & Beachy, 1989). Different movement proteins produced by viruses share few common features like binding to nucleic acid (Citovsky, Knorr, Schuster, & Zambryski, 1990), specific localization at plasmodesmata (Ding et al., 1992) and increasing the exclusion size limit of plasmodesmata (Wolf et al., 1989). Similar to the viral replication, chloroplast and associated factors also play an important role in intercellular and systemic virus movement inside the host plant. After infection of *Alternanthera mosaic virus* (AltMV), expression of triple gene block-3 (TGB-3) gene leads to the preferential accumulation of its protein in the mesophyll cells after receiving information from chloroplast after having chloroplast targeted signal which is very important for the virus movement in plant (Lim et al., 2010). Similarly, MP of *Abutilon mosaic virus* (AbMV) interacts with cpHSC70-1, a chloroplast targeted heat shock protein and subsequently colocalize inside the chloroplast (Krenz, Jeske, & Kleinow, 2012). So, it was commonly observed that during virus-chloroplast interactions, viral factor initially interacts with chloroplast factor and then hijack them and use them for their movement. Several viruses like *Cauliflower mosaic virus* (CaMV), TMV and *Tomato mosaic virus* (ToMV) were reported to target chloroplast and their

factors for the efficient intercellular and systemic movement of these viruses inside the host plant (Angel et al., 2013; Hohn, Fütterer, & Hull, 1997; Rodriguez et al., 2014; Zhang et al., 2008; Zhao et al., 2013).

26.4.1 Changes in chloroplast structure upon viral infection

To invade the host cell environment to an optimal level, numerous biochemical, and physiological features of host cell are utilized by the viruses upon infection, such as disruption of double membrane structure of chloroplasts. These rearrangements and alterations in the ultrastructure organization of the plant cell have been reported by various research studies (Allen, 1972; Appiano, Pennazio, & Redolfi, 1978; Arnott, Rosso, & Smith, 1969; Bald, 1948; Laliberte & Sanfacon, 2010; Li et al., 2006; Musetti, Bruni, & Favali, 2002; Yan, Lehrer, Hajirezaei, Springer, & Komor, 2008; Zhao et al., 2016). Virus infection is known to hamper numerous genes involved in chloroplast and photosynthesis activity and includes chlorophyll synthesis enzymes, antenna proteins, chlorophyll catabolism, Electron transport chain, Antioxidant defense component of chloroplasts, chloroplasts differentiation, PSI and PSII related proteins, Rubisco proteins etc. Further, reported evidences also highlight the manipulation of protein sorting mechanism in plant cells by viruses through molecular mimicry. For instance, Coat Protein (CP) of Cucumber necrosis virus (CNV) localize itself to the thylakoid membrane as it possess N-terminal domain bearing sequence similarity to transit peptide (TPs) of chloroplast (Xiang et al., 2006). However, CP of TMV, Potato virus X (PVX), chloroplast targeting of P5-2 of Rice black-streaked dwarf virus (Bhattacharyya et al., 2015) deploy other mechanisms to make their way to chloroplast. CP of PVX has been demonstrated not to be synthesized from viral RNA in the chloroplast; rather its posttranslational localization to chloroplast following its interaction with the TP of the plastocyanin precursor has been revealed by Qiao, Li, Wong, and Fan (2009). Barley stripe mosaic virus (BSMV) causes the structural and functional retardation of plastids via altering the lipid composition of etioplasts, thus affecting the translocation of NADPH: protochlorophyllide oxidoreductase, a nuclear-encoded plastid inner membrane protein. The lipid-facilitated binding of precursor proteins on chloroplasts surface is imperative as it governs the translocation of nuclear encoded proteins into plastids (Harsanyi et al., 2006). Though many research studies have been conducted, albeit still detailed insight are required to decode the precise mechanisms adopted by the viruses to conquer the chloroplasts upon infection.

Some of the primary changes that take place in the chloroplast ultrastructure during viral infection includes:

1. Chloroplast clustering and a decline in their overall number.
2. Atypical chloroplast appearances such as presence of swollen or globule type chloroplasts, or some chloroplasts becomes amoeboid shaped with membrane bounded extrusion and in some cases there would be the generation of stromules.
3. Normal structure of peripheral vesicle and cytoplasm invagination get disturbed, membrane proliferations takes place and envelope get broken.
4. Size of chloroplast vacuoles and vesicles get reduced whereas, size of intermembranous sacs get increased, large number of enlarged starch grain get synthesized, size and number of electron dense plastoglobules/granules/bodies get increased.
5. Disappearance of grana stacks, thylakoids get distorted/dilated/loosen
6. Entire chloroplast gets destroyed and disorganized and grana gets scattered in cytoplasm.

26.4.2 Virus factors involved in structural and functional changes of chloroplast

The ultimate fate of chloroplast-virus interaction is governed by maintaining the dynamic equilibrium between activation of plant defense response and sequestering of chloroplast proteins such as PsbP by viruses that facilitates effective viral replication. The implications of these interactions largely depend on the localization of host proteins and could be mediated either in the thylakoid membrane or lumen, stroma, chloroplast membrane or cytosol.

Earlier many workers reported that formation of different virus inclusion bodies or virion like particles in chloroplast is directly related to the development of viral infection symptoms (Shalla, 1964; Zhao et al., 2013). It was also reported that these virions like particles are actually pseudovirion in nature where transcripts of chloroplast are encapsidated by coat protein of TMV (Atreya & Siegel, 1989; Shalla, Petersen, & Giunchedi, 1975) which showed the involvement of coat protein in chloroplast ultrastructure alteration (Banerjee & Zaitlin, 1992). Coat protein of different viruses affects the different components of chloroplast. In TMV, coat protein affects the thylakoid membrane component of chloroplast in artificially infected tobacco leaves (Hodgson, Beachy, & Pakrasi, 1989; Reinero & Beachy, 1986) and found to induce ultrastructure rearrangements in the chloroplast. Infection by Tobamovirus causes its coat protein to affect the

chloroplast structure by binding with ferredoxin-I component (Sun et al., 2013). Coat protein of *Potato virus X* was reported to affect the grana stacks and membranes of chloroplast (Kozar & Sheludko, 1969; Qiao et al., 2009). However, in *Potato virus Y* (PVY), coat protein was reported to affect the membranes of thylakoid (Gunasinghe & Berger, 1991). Along with effecting thylakoid membrane, coat protein of *Potato virus Y* was also found to affect the large subunits of RuBisCo (RbcL) that leads to the chlorosis development and mosaic symptoms (Feki et al., 2005). CMV infection also leads to the change in chloroplast ultrastructure (Mazidah, Lau, Yusoff, Habibuddin, & Tan, 2012; Roberts & Wood, 1982). Coat protein of CMV can get transported inside the intact chloroplast in an Adenosine Tri Phosphate (ATP)-independent mode and severity of mosaic symptoms is directly correlated with the amount of coat protein present in chloroplast (Liang, Ye, Shi, Kang, & Tian, 1998). In addition to coat protein, there are some other factors also present in viruses that cause alteration in chloroplast ultra structure. Unlike *Potato virus X* TGB3, *Potexvirus* AltMV TGB-3 has a chloroplast targeting signal and accumulates around membrane of chloroplast preferentially (Lim et al., 2010). However, vesiculation of the chloroplast membrane and development of venial necrosis symptoms also take place due to the over expression of AltMv TGB3 (Jang et al., 2013; Lim et al., 2010). TGB-3 protein of AltMv was reported to interact with PsbO [PSII oxygen evolving complex (OEC) protein] and this interaction was found to be very crucial in the development of virus infection symptoms and disruption of chloroplast (Jang et al., 2013). In case of PVY infection, change in size and number of chloroplast takes place as helper component - proteinase (HC-Pro) protein that is a viral multifunctional protein interact with Min D factor that is chloroplast division related factor (Lin, Ding, Hsu, & Tsai, 2007). Some of these interactions between the proteins of virus and chloroplasts and their implications are listed in Table 26.2.

Several transcriptomic and proteomic studies have established disruption of numerous molecular events in plants following virus infection. Majority of chloroplast proteins get affected during virus infection and most of them get down regulated and are directly correlated with the chlorosis severity (Dardick, 2007; Mochizuki et al., 2014; Rodriguez, Munoz, Lenardon, & Lascano, 2012; Wu et al., 2013). Chloroplast photosynthesis related proteins (CPRPs) are the common target during viral infection. Out of different CPRPs, RbCs and Rubisco activase in the stroma of chloroplast whereas, OEC and light harvesting antenna complex in the PSII of thylakoid get majorly effected (Kundu, Chakraborty, Kundu, & Pal, 2013; Liu, Yang, Bi, & Zhang, 2014; Moshe et al., 2012; Pineda, Sajjani, & Baron, 2010; Wang, Hajano, Ren, Lu, & Wang, 2015). Plant pathogenic viruses can affect the biosynthesis of CPRPs at various stages that include their transcription, translation, posttranscriptional, chloroplast transportation, assembly, and/or their degradation in the plastid that ultimately contributes towards the symptom development (Lehto, Tikkanen, Hiriart, Paakkari, & Aro, 2003; Perez-Bueno, Rahoutei, Sajjani, Garcia-Luque, & Baron, 2004).

26.5 Role of chloroplast in the defense against plant pathogenic viruses

Chloroplast induced defense action in response to pathogen requires the elicitor signal molecules to localize in the chloroplast or presence of receptors on chloroplast membrane which in turn generate a retrograde signal from chloroplast to the nuclei for subsequent activation of defense components. For instance, Toll interleukin receptor (TIR)-NB-LRRs

TABLE 26.2 Interaction between chloroplast–virus proteins and their implications.

Chloroplast–virus interaction	Virus (gene)	Chloroplast protein
Hampering of host protein translocation into chloroplast	Soybean Mosaic Virus (P1) Sugarcane Mosaic Virus (HC-Pro)	Rieske Fe/S Ferredoxin-5 precursor
Regulation of defense components	Tomato Mosaic Virus (MP) Tobacco Mosaic Virus (Replicase)	RbcS ATP synthase γ -subunit (AtpC), RuBisCO activase (RCA)
Regulation of virus infection	Plum Pox Virus (CI) Potato Virus X (CP)	Photosystem I PSI-K RuBisCO large subunit
Affects Chloroplast stability	Tomato Mosaic Virus (CP) Rice Stripe Virus (Disease-specific protein)	IP-L PsbP
ETI elicitation	Tobacco Mosaic Virus (Replicase)	NRIP 1

possesses putative chloroplast localization signals. Further, receptors which are not localized in the organelle, recognize pathogen through interaction with chloroplast proteins. For example, in TMV infection, helicase domain of TMV replicase (p50) is recognized by TIR domain belonging to TIR-NB-LRRs class through a chloroplast localized protein N receptor-interacting protein 1 (NRIP1), which gets released from chloroplast in the presence of p50, to form a tripartite complex that subsequently activates defense signaling (Caplan, Mamillapalli, Burch-Smith, Czymmek, & Dinesh-Kumar, 2008). TMV p50 induced, N-mediated defense activation results in the formation of stromules (array of tubular structures filled with stroma) from chloroplast to nucleus and elicits HR-programmed cell death (PCD) (Caplan et al., 2015).

Various aspects of chloroplast mediated defense action are described below:

1. Chloroplasts are rich source of reactive oxygen species (ROS), which could elicit HR/effector—triggered immunity and ultimately PCD in the plant
2. Synthesis of defense hormones in local and systemic defense responses, primarily salicylic acid (SA) and along with jasmonic acid (JA) and abscisic acid (ABA) in regulation by chloroplast apparatus (Caplan et al., 2008)
3. Level of calcium (Ca^{2+}) pool stored inside chloroplasts changes in response to immune response and pathogen invading (Mur, Kenton, Lloyd, Ougham, & Prats, 2008)
4. Since gene silencing machinery is not available in the chloroplasts, cross-talk between chloroplast derived signaling molecules and RNA silencing is of utmost validity.
5. Light dependent regulation of expression of small subunit of RuBisCO; interaction between RbcS with other components such as OEC to stimulate defense actions

Chloroplast is the major site for the salicylic acid (SA) and jasmonic acid (JA) biosynthesis in providing resistance to the plants against biotrophic plant pathogens by regulating systemic acquired immunity (Lin et al., 2015; Wasternack & Hause, 2013; Wasternack, 2007). For instance, SA promotes plant defense response against viruses and its biosynthesis (through its exogenous application/or its analogs) and signaling in the resistant plant varieties get enhanced and accounts for basal immunity (Falcioni et al., 2014; Garcion et al., 2008; Wildermuth, Dewdney, Wu, & Ausubel, 2001). Along with these hormone biosynthesis, there are some factors of chloroplast that controls antagonistic interaction of SA-JA synthesis and signaling (Kunkel & Brooks, 2002; Lemos et al., 2016; Zheng et al., 2012). Further, expression of chloroplast's calcium sensing receptors is reported to be increased with SA accumulation that links cytoplasmic-nuclear immune responses to the chloroplast (Nomura et al., 2012). Similar to salicylic acid, JA also plays an important role in controlling disease during compatible plant–virus interactions (Alazem & Lin, 2015).

Further, chloroplast is also an important site for the synthesis of ROS that plays an important role in the plant defense during incompatible plant–virus interactions (Allan, Lapidot, Culver, & Fluhr, 2001; Hakmaoui et al., 2012). ROS burst is an important component of HR against various plant pathogens during incompatible interaction (Torres, Jones, & Dangl, 2006; Zurbriggen, Carrillo, & Hajirezaei, 2010). Moreover, retrograde signaling from chloroplasts to nucleus to activate immune response is pivotal in resistance response. Particularly, stromules of chloroplast are involved in magnifying and transporting the defense related signals to the nucleus. For instance, TMV infection has been observed to bring about the accumulation of chloroplast localized defense components NRIP1 and H_2O_2 in the nucleus (Zhao et al., 2016). Lastly, changes in the chloroplast ultrastructure, such as TMV infection causes chloroplast swelling and bursting of its membrane during the N-mediated hypersensitive reaction results in the resistance response.

26.6 Plant–virus metagenomics

To further expand the avenues of plant–virus biodiversity, metagenomics analysis is the new approach for identifying causative viral disease features, screening of viruses, detecting novel, cryptic, or asymptomatic viruses, recognizing new viral strains/species/families, identifying virus agents in a single or complex infections, etc., through sequencing of viral populations present in a particular environmental sample (MacDiarmid, Rodoni, Melcher, Ochoa-Corona, & Roossinck, 2013; Roossinck, Martin, & Roumagnac, 2015). From a particular sample, several technologies exist to enrich for the plant viral specific sequences such as dsRNA enrichment (Roossinck et al., 2010), siRNA (Kreuze et al., 2009), or isolation of virus like particles (Muthukumar et al., 2009) to provide a deeper insight into host specific- or environment specific virus–plant interactions. As Plant–virus interaction in terms of appearance of disease symptoms are widely studied, further, one could also explore the paths of characterizing plant–virus collaboration as conditional mutualists, cross-protection agents etc. (Fraser, 1998; Roossinck, 2011) in special reference to plastome specific or plastome directed changes. Wamonje et al. (2017) utilized NGS to explore dicistrovirus diversity in maize and their insect vectors, aphids and identified strains of a novel Big Sioux River virus (BSRV) -like, along with Rhopalosiphum padi

virus (RhPV), Aphid lethal paralysis virus (ALPV) dicistrovirus in *Aphis fabae* and maize. These viruses are known to use the plants in which they do not replicate as a reservoir to infect new insect hosts wherein the virus could replicate. Moreover, viruses rarely causes any visible symptom upon infection in their wild host species and chloroplast genome structural features of these wild hosts needs to be explored in a larger details to better capture the plant–virus ecology relative to the cultivated counterparts.

26.7 Conclusion

Advent of novel molecular tools and techniques has provided a deeper insight into chloroplast–virus interaction studies, though still a nano-scale investigation in terms of chloroplast genes, virus proteins, alleles, their interactions, implications, manipulations etc. needs to be carried out. Retrograde signaling mediated by chloroplasts is another promising aspect for further exploration. Chloroplast genome sequence of numerous plant species have been deciphered and could be utilized for the plastome engineering strategies, genome editing tools etc. for boosting viral resistance. Numerous sequence databases and bioinformatics tools have aided in the characterization of virus and their interaction with their host through *in silico* studies. Though this still requires precise pinpointing of the candidate chloroplast genes/proteins involved and decoding of their metabolic pathways. Viral metagenomics has also opened new avenues to be discerned to aid in development of virus-resistant plants. By studying a careful interaction between virus and its wild and cultivated host, differences, and similarities between genes, pathways, mechanisms, etc., could be interpreted. Thus chloroplast–virus interaction studies require a multidisciplinary approach by molecular biologists, virologists, bioinformaticians, and geneticists, etc., to contribute towards the future development of plants with virus resistance.

References

- Ahlquist, P., Noueir, A. O., Lee, W. M., Kushner, D. B., & Dye, B. T. (2003). Host factors in positive-strand RNA virus genome replication. *Journal of Virology*, *77*(15), 8181–8186.
- Alazem, M., & Lin, N. S. (2015). Roles of plant hormones in the regulation of host–virus interactions. *Molecular Plant Pathology*, *16*(5), 529–540.
- Allan, A. C., Lapidot, M., Culver, J. N., & Fluhr, R. (2001). An early tobacco mosaic virus-induced oxidative burst in tobacco indicates extracellular perception of the virus coat protein. *Plant Physiology*, *126*(1), 97–108.
- Allen, T. C. (1972). Subcellular responses of mesophyll cells to wild cucumber mosaic virus. *Virology*, *47*(2), 467–474.
- Almasi, A., Harsanyi, A., & Gaborjanyi, R. (2001). Photosynthetic alterations of virus infected plants. *Acta Phytopathologica et Entomologica Hungarica*, *36*, 15–29.
- Angel, C. A., Lutz, L., Yang, X., Rodriguez, A., Adair, A., Zhang, Y., et al. (2013). The P6 protein of cauliflower mosaic virus interacts with CHUP1, a plant protein which moves chloroplasts on actin microfilaments. *Virology*, *443*(2), 363–374.
- Appiano, A., Pennazio, S., & Redolfi, P. (1978). Cytological alterations in tissues of *Gomphrena globosa* plants systemically infected with tomato bushy stunt virus. *Journal of General Virology*, *40*, 277–286.
- Arnott, H. J., Rosso, S. W., & Smith, K. M. (1969). Modification of plastid ultrastructure in tomato leaf cells infected with tobacco mosaic virus. *Journal of Ultrastructure Research*, *27*(2), 149–167.
- Asaf, S. K., Khan, A. L., Shahzad, M. A., Lubna, R., Kang, S. M., et al. (2018). Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. *PLoS One*, *13*(3), e0192966.
- Asurmendi, S., Berg, R. H., Koo, J. C., & Beachy, R. (2004). Coat protein regulates formation of replication complexes during tobacco mosaic virus infection. *Proceedings of the National Academy of Science*, *101*(5), 1415–1420.
- Atreya, C. D., & Siegel, A. (1989). Localization of multiple TMV encapsidation initiation sites on rbcL gene transcripts. *Virology*, *168*(2), 388–392.
- Balachandran, S., Osmond, C. B., & Daley, P. F. (1994). Diagnosis of the earliest strain-specific interactions between tobacco mosaic virus and chloroplasts of tobacco leaves *in vivo* by means of chlorophyll fluorescence imaging. *Plant Physiology*, *104*, 1059–1065.
- Bald, J. G. (1948). The development of amoeboid inclusion bodies of tobacco mosaic virus. *Australian Journal of Botany*, *1*, 458–463.
- Banerjee, N., & Zaitlin, M. (1992). Import of tobacco mosaic virus coat protein into intact chloroplasts *in vitro*. *Molecular Plant-Microbe Interactions*, *5*, 466–471.
- Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, *6*(1), 106–136.
- Bhat, S., Folimonova, S. Y., Cole, A. B., Ballard, K. D., Lei, Z., Watson, B. S., et al. (2012). Influence of host chloroplast proteins on Tobacco mosaic virus accumulation and intercellular movement. *Plant Physiology*, *161*, 134–147.
- Bhattacharyya, D., Prabu, G., Reddy, K. K., Kushwaha, N. K., Sharma, V. K., Yusuf, M. A., et al. (2015). A geminivirus betasatellite damages the structural and functional integrity of chloroplasts leading to symptom formation and inhibition of photosynthesis. *Journal of Experimental Botany*, *66*, 5881–5895.
- Blawid, R., Silva, J. M. F., & Nagata, T. (2017). Discovering and sequencing new plant viral genomes by next-generation sequencing: Description of a practical pipeline. *Annals of Applied Biology*, *170*(3), 301–314.

- Bock, R. (2007). Plastid biotechnology: Prospects for herbicide and insect resistance, metabolic engineering, and molecular farming. *Current Opinions in Biotechnology*, 18, 100–106.
- Bock, R., & Warzecha, H. (2010). Solar-powered factories for new vaccines and antibiotics. *Trends in Biotechnology*, 28, 246–252.
- Bruneau, A., Doyle, J. J., & Palmer, J. D. (1990). A Chloroplast DNA inversion as a subtribal character in the Phaseoleae (Leguminosae). *Systematic Botany*, 15, 378–386.
- Budziszewska, M., & Steplowska, O. (2018). The role of chloroplast in the replication of positive-sense single stranded plant RNA viruses. *Frontiers in Microbiology*, 9, 1776.
- Cai, Z., Guisinger, M., Kim, H. G., Ruck, E., Blazier, J. C., McMurtry, V., et al. (2008). Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *Journal of Molecular Evolution*, 67, 696–704.
- Caplan, J. L., Kumar, A. S., Park, E., Padmanabhan, M. S., Hoban, K., Modla, S., et al. (2015). Chloroplast stromules function during innate immunity. *Developmental Cell*, 34, 45–57.
- Caplan, J. L., Mamillapalli, P., Burch-Smith, T. M., Czymbek, K., & Dinesh-Kumar, S. P. (2008). Chloroplastic protein NRIP1 mediates innate immune receptor recognition of a viral effector. *Cell*, 132, 449–462.
- Chen, I., Tsai, A. Y., Huang, Y. P., Wu, I. F., Cheng, S. F., Hsu, Y. H., et al. (2017). Nuclear encoded plastidal carbonic anhydrase is involved in replication of Bamboo mosaic virus RNA in *Nicotiana benthamiana*. *Frontiers in Microbiology*, 8, 2046.
- Chen, J., Hao, Z., Xu, H., Yang, L., Liu, G., Sheng, Y., et al. (2015). The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides*. *Frontiers in Plant Sciences*, 6, 447.
- Chen, T., Wang, X., Von Wangenheim, D., Zheng, M., Samaj, J., Ji, W., & Lin, J. (2012). Probing and tracking organelles in living plant cells. *Protoplasma*, 249(2), 157–167.
- Cheng, H., Li, J., Zhang, H., Cai, B., Gao, Z., Qiao, Y., et al. (2017). The complete chloroplast genome sequence of strawberry (*Fragaria ananassa* Duch.) and comparison with related species of Rosaceae. *Peer the Journal*, 5, e3919.
- Cheng, J., Zeng, X., Ren, G., & Liu, Z. (2013). CGAP: A new comprehensive platform for the comparative analysis of chloroplast genomes. *BMC Bioinformatics*, 14.
- Choudhury, S., Larkin, P., Meinke, H., Hasanuzzaman, M. D., Johnson, P., & Zhou, M. (2019). Barley yellow dwarf virus infection affects physiology, morphology, grain yield and flour pasting properties of wheat. *Crop & Pasture Science*, 70(1), 16–25.
- Citovsky, V., Knorr, D., Schuster, G., & Zambryski, P. (1990). The P30 movement protein of tobacco mosaic virus is a single-strand nucleic acid binding protein. *Cell*, 60(4), 637–647.
- Clarke, J. L., & Daniell, H. (2011). Plastid biotechnology for crop production: Present status and future perspectives. *Plant Molecular Biology*, 76, 211–220.
- Cowan, G. H., Roberts, A. G., Chapman, S. N., Ziegler, A., Savenkov, E. I., & Torrance, L. (2012). The potato mop-top virus TGB2 protein and viral RNA associate with chloroplasts and viral infection induces inclusions in the plastids. *Frontiers in Plant Science*, 3, 290.
- Curci, P. L., De Paola, D., Danzi, D., Vendramin, G. G., & Sonnante, G. (2015). Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. *PLoS One*, 10, e0120589.
- D'Agostino, N., Tamburino, R., Cantarella, C., De Carluccio, V., Sannino, L., Cozzolino, S., et al. (2018). The complete plastome sequences of eleven *Capsicum* genotypes: Insights into DNA variation and molecular evolution. *Genes*, 9, 503.
- Daniell, H., Choun, S. L., Ming, Y., & Wan, J. C. (2016). Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biology*, 17, 134.
- Daniell, H., Singh, N. D., Mason, H., & Streatfield, S. J. (2009). Plant-made vaccine antigens and biopharmaceuticals. *Trends in Plant Science*, 14, 669–679.
- Daniell, H., Wurdack, K. J., Kanagaraj, A., Lee, S. B., Sasaki, C., & Jansen, R. K. (2008). The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. *Theoretical and Applied Genetics*, 116, 723–737.
- Dardick, C. (2007). Comparative expression profiling of *Nicotiana benthamiana* leaves systemically infected with three fruit tree viruses. *Molecular Plant Microbe Interactions*, 20(8), 1004–1017.
- Das, P. P., Lin, Q., & Wong, S. M. (2018). Comparative proteomics of Tobacco mosaic virus-infected *Nicotiana tabacum* plants identified major host proteins involved in photosystems and plant defence. *Journal of Proteomics*, 1(194), 191–199.
- De Cosa, B., Moar, W., Lee, S. B., Miller, M., & Daniell, H. (2001). Overexpression of the *Bt* Cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. *Nature Biotechnology*, 19, 71–74.
- De Graaff, M., Coscoy, L., & Jaspars, E. M. J. (1993). Localization and biochemical characterization of Alfalfa mosaic virus replication complexes. *Virology*, 194, 878–881.
- Dierckxsens, N., Mardulyn, P., & Smits, G. (2016). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*. Available from <https://doi.org/10.1093/nar/gkw955>.
- Ding, B., Haudenschild, J. S., Hull, R. J., Wolf, S., Beachy, R. N., & Lucas, W. J. (1992). Secondary plasmodesmata are specific sites of localization of the tobacco mosaic virus movement protein in transgenic tobacco plants. *The Plant Cell*, 4, 915–928.
- Dong, W., Liu, J., Yu, J., Wang, L., & Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA Barcoding. *PLoS One*, 7, e35071.

- Doyle, J. J., Doyle, J. L., & Palmer, J. D. (1995). Multiple independent losses of two genes and one intron from legume chloroplast genomes. *Systematic Botany*, 20, 272–294.
- Dreher, T. W. (2004). Turnip yellow mosaic virus: Transfer RNA mimicry, chloroplasts and a C-rich genome. *Molecular Plant Pathology*, 5(5), 367–375.
- Falcioni, T., Ferrio, J. P., Del Cueto, A. I., Gine, J., Achon, M. A., & Medina, V. (2014). Effect of salicylic acid treatment on tomato plant physiology and tolerance to potato virus X infection. *European Journal of Plant Pathology*, 138, 331–345.
- Feki, S., Loukili, M. J., Triki-Marrakchi, R., Karimova, G., Old, I., Ounouna, H., et al. (2005). Interaction between tobacco ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (RubisCO-LSU) and the PVY coat protein (PVY-CP). *European Journal of Plant Pathology*, 112, 221–234.
- Fraser, R. S. S. (1998). Introduction to classical cross protection. In G. D. Foster, & S. C. Taylor (Eds.), *Plant virology protocols* (pp. 13–24). Totowa, NJ: Humana Press. Available from <http://doi.org/10.1385/0-89603-385-6>.
- Garcion, C., Lohmann, A., Lamodièrè, E., Catinot, J., Buchala, A., Doermann, P., et al. (2008). Characterization and biological function of the *ISOCHORISMATE SYNTHASE2* gene of *Arabidopsis*. *Plant Physiology*, 147, 1279–1287.
- Gunasinghe, U., & Berger, P. (1991). Association of potato virus Y gene products with chloroplasts in tobacco. *Molecular Plant Microbial Interaction*, 4, 452–457.
- Guo, X., Castillo-Ramírez, S., González, V., Bustos, P., Fernández-Vázquez, J. L., Santamaría, R. I., et al. (2007). Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and the genomic diversification of legume chloroplasts. *BMC Genomics*, 8, 228.
- Hadidi, A., Flores, R., Candresse, T., & Barba, M. (2016). Next-generation sequencing and genome editing in plant virology. *Frontiers in Microbiology*, 7–19.
- Hagemann, R. (2004). The sexual inheritance of plant organelles. In H. Daniell, & C. Chase (Eds.), *Molecular biology and biotechnology of plant organelles* (pp. 87–108). Dordrecht, The Netherlands: Springer.
- Hakmaoui, A., Perez-Bueno, M. L., Garcia-Fontana, B., Camejo, D., Jimenez, A., Sevilla, F., et al. (2012). Analysis of the antioxidant response of *Nicotiana benthamiana* to infection with two strains of pepper mild mottle virus. *Journal of Experimental Botany*, 63(15), 5487–5496.
- Han, Z., Li, W., Liu, Y., & Gao, L. (2016). The complete chloroplast genome of North American ginseng, *Panax quinquefolius*. *Mitochondrial DNA*, 27, 3496–3497.
- Harsanyi, A., Ryberg, M., Andersson, M. X., Boka, K., Laszlo, L., Botond, G., et al. (2006). Alterations of NADPH: protochlorophyllide oxidoreductase quantity and lipid composition in etiolated barley seedlings infected by Barley stripe mosaic virus (BSMV). *Molecular Plant Pathology*, 7, 533–541.
- Herbers, K., Takahata, Y., Melzer, M., Mock, H. P., Hajirezaei, M., & Sonnwald, U. (2000). Regulation of carbohydrate partitioning during the interaction of potato virus Y with tobacco. *Molecular Plant Pathology*, 1, 51–59.
- Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., et al. (1989). The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Molecular Genetics and Genomics*, 217, 185–194.
- Hodgson, R. A., Beachy, R. N., & Pakrasi, H. B. (1989). Selective inhibition of photosystem II in spinach by tobacco mosaic virus: An effect of the viral coat protein. *FEBS Letters*, 245, 267–270.
- Hohn, T., Fütterer, J., & Hull, R. (1997). The proteins and functions of plant pararetroviruses: Knowns and unknowns. *Critical Review in Plant Science*, 16, 133–161.
- Huang, D. I., & Cronk, Q. C. (2015). Plann: A command-line application for annotating plastome sequences. *Applications in Plant Sciences*, 3(8). Available from <https://doi.org/10.3732/apps.1500026>.
- Huang, Y. P., Chen, I., & Tsai, C. H. (2017). Host factors in the infection cycle of *Bamboo mosaic virus*. *Frontiers in Microbiology*, 8, 437.
- Hull, R. (2014). Plant viruses and their classification. In R. Hull (Ed.), *Plant virology* (pp. 15–68). Ciudad: Academic Press.
- Jakubiec, A., Notaise, J., Tournier, V., Hericourt, F., Block, M. A., Drugeon, G., et al. (2004). Assembly of turnip yellow mosaic virus replication complexes: Interaction between the proteinase and polymerase domains of the replication proteins. *Journal of Virology*, 78(15), 7945–7957.
- Jang, C., Seo, E. Y., Nam, J., Bae, H., Gim, Y. G., Kim, H. G., et al. (2013). Insights into *Alternanthera mosaic virus* TGB3 functions: Interactions with *Nicotiana benthamiana* PsbO correlate with chloroplast vesiculation and veinal necrosis caused by TGB3 over-expression. *Frontiers in Plant Science*, 4, 5.
- Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Depamphilis, C. W., Leebens-Mack, J., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 19369–19374.
- Jansen, R. K., Sasaki, C., Lee, S. B., Hansen, A. K., & Daniell, H. (2011). Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for at least two independent transfers of rpl22 to the nucleus. *Molecular Biology and Evolution*, 28, 835–847.
- Jansen, R. K., Wojciechowski, M. F., Sanniyasi, E., Lee, S. B., & Daniell, H. (2008). Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). *Molecular Phylogenetics and Evolution*, 48, 1204–1217.
- Jarvis, P., & Soll, J. (2001). Toc, tic, and chloroplast protein import. *Biochimica et Biophysica Acta*, 1541, 64–79.
- Jeong, Y. M., Chung, W. H., Mun, J. H., Kim, N., & Yu, H. J. (2014). *De novo* assembly and characterization of the complete chloroplast genome of radish (*Raphanus sativus* L.). *Gene*, 551(1), 39–48.
- Jin, S., & Daniell, H. (2015). Engineered chloroplast genome just got smarter. *Trends in Plant Science*, 20(10), 622–640.

- Kaila, T., Chaduvla, P. K., Rawal, H. C., Saxena, S., Tyagi, A., Mithra, S., et al. (2017). Chloroplast genome sequence of clusterbean (*Cyamopsis tetragonoloba* L.): Genome structure and comparative analysis. *Genes*, 8(9), 212.
- Kaila, T., Chaduvla, P. K., Saxena, S., Bahadur, K., Gahukar, S. J., Chaudhury, A., et al. (2016). Chloroplast genome sequence of pigeonpea (*Cajanus cajan* (L.) Millspaugh) and *Cajanus scarabaeoides* (L.) Thouars: Genome organization and comparison with other legumes. *Frontiers in Plant Science* (7, p. 1847).
- Kato, T., Kaneko, T., Sato, S., Nakamura, Y., & Tabata, S. (2000). Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Research*, 7, 323–330.
- Keller, J., Rousseau-Gueutin, M., Martin, G. E., Morice, J., Boutte, J., Coissac, E., et al. (2017). The evolutionary fate of the chloroplast and nuclear rps16 genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from *Lupinus*. *DNA Research*, 24, 343–358.
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research*, 27(5), 722–736.
- Kozar, F. E., & Sheludko, Y. M. (1969). Ultrastructure of potato and *Datura stramonium* plant cells infected with potato virus X. *Virology*, 38(2), 220–229.
- Krenz, B., Jeske, H., & Kleinow, T. (2012). The induction of stromule formation by a plant DNA-virus in epidermal leaf tissues suggests a novel intra- and intercellular macromolecular trafficking route. *Frontiers in Plant Science*, 3, 291.
- Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., et al. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology*, 388, 1–7.
- Kundu, S., Chakraborty, D., Kundu, A., & Pal, A. (2013). Proteomics approach combined with biochemical attributes to elucidate compatible and incompatible plant-virus interactions between *Vigna mungo* and mungbean yellow mosaic India virus. *Proteome Science*, 11, 15.
- Kunkel, B. N., & Brooks, D. M. (2002). Cross talk between signaling pathways in pathogen defense. *Current Opinion in Plant Biology*, 5(4), 325–331.
- Kyselakova, H., Prokopova, J., Naus, J., Novak, O., Navratil, M., Safarova, D., et al. (2011). Photosynthetic alterations of pea leaves infected systemically by pea enation mosaic virus: A coordinated decrease in efficiencies of CO₂ assimilation and photosystem II photochemistry. *Plant Physiology and Biochemistry*, 49(11), 1279–1289.
- Laliberte, J. F., & Sanfacon, H. (2010). Cellular remodeling during plant virus infection. *Annual Review of Phytopathology*, 48, 69–91.
- Lazarowitz, S. G., & Beachy, R. N. (1999). Viral movement proteins as probes for intracellular and intercellular trafficking in plants. *The Plant Cell*, 11, 535–548.
- Lee, S. B., Kaittanis, C., Jansen, R. K., Hostetler, J. B., Tallon, L. J., Town, C. D., & Daniell, H. (2006). The complete chloroplast genome sequence of *Gossypium hirsutum*: Organization and phylogenetic relationships to other angiosperms. *BMC Genomics*, 7, 61.
- Lehto, K., Tikkanen, M., Hiriart, J. B., Paakkari, V., & Aro, E. M. (2003). Depletion of the photosystem II core complex in mature tobacco leaves infected by the flavum strain of tobacco mosaic virus. *Molecular Plant-Microbe Interaction*, 16(12), 1135–1144.
- Leister, D. (2003). Chloroplast research in the genomic age. *Trends in Genetics*, 19, 47–56.
- Lemos, M., Xiao, Y., Bjornson, M., Wang, J. Z., Hicks, D., Souza, A. D., et al. (2016). The plastidial retrograde signal methyl erythritol cycloprophosphate is a regulator of salicylic acid and jasmonic acid crosstalk. *Journal of Experimental Botany*, 67(5), 1557–1566.
- Lenz, H., & Knoop, V. (2013). PREPACT 2.0: Predicting C-to-U and U-to-C RNA editing in organelle genome sequences with multiple references and curated RNA editing annotation. *Bioinformatics and Biology Insights*, 7.
- Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: From gene to genome. *Biological Reviews*, 90, 157–166.
- Li, Y., Cui, H., Cui, X., & Wang, A. (2016). The altered photosynthetic machinery during compatible virus infection. *Current Opinions in Virology*, 17, 19–24.
- Li, Y. H., Hong, J., Xue, L., Yang, Y., Zhou, X. P., & Jiang, D. A. (2006). Effects of Broad bean wilt virus 2 isolate infection on photosynthetic activities and chloroplast ultrastructure in broad bean leaves. *Journal of Plant Physiology and Molecular Biology*, 32(4), 490–496.
- Liang, D., Ye, Y., Shi, D., Kang, L., & Tian, B. (1998). The role of viral coat protein in the induction of mosaic symptoms in tobacco. *Scientia Sinica Vitae*, 28, 251–256.
- Lim, H. S., Vaira, A. M., Bae, H., Bragg, J. N., Ruzin, S. E., Bauchan, G. R., et al. (2010). Mutation of a chloroplast-targeting signal in alternanthera mosaic virus TGB3 impairs cell-to cell movement and eliminates long-distance virus movement. *Journal of General Virology*, 91(8), 2102–2115.
- Lin, C. S., Chen, J., Huang, Y. T., Chan, M. T., Daniell, H., Chang, W. J., et al. (2015). The location and translocation of *ndh* genes of chloroplast origin in the Orchidaceae family. *Scientific Reports*, 5, 9040.
- Lin, J. W., Ding, M. P., Hsu, Y. H., & Tsai, C. H. (2007). Chloroplast phosphoglycerate kinase, a gluconeogenic enzyme, is required for efficient accumulation of bamboo mosaic virus. *Nucleic Acids Research*, 35(2), 424–432.
- Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, 13, 715. Available from <https://doi.org/10.1186/1471-2164-13-715>.
- Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., et al. (2010). Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *Journal of Virology*, 84, 11879–11887.
- Liu, J., Yang, J., Bi, H., & Zhang, P. (2014). Why mosaic? Gene expression profiling of African cassava mosaic virus-infected cassava reveals the effect of chlorophyll degradation on symptom development. *Journal of Integrative Plant Biology*, 56(2), 122–132.

- Lohse, M., Drechsel, O., & Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics*, 52(5–6), 267–274.
- Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). OrganellarGenomeDRAW—A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research*, 289.
- MacDiarmid, R., Rodoni, B., Melcher, U., Ochoa-Corona, F., & Roossinck, M. (2013). Biosecurity implications of new technology and discovery in plant virus research. *PLoS Pathogens*, 9, e1003337.
- Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Sémon, M., Perry, A. S., et al. (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research*, 20, 1700–1710.
- Maier, R. M., Neckermann, K., Igloi, G. L., & Kössel, H. (1995). Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *Journal of Molecular Biology*, 251, 614–628.
- Mandahar, C. L., & Garg, I. D. (1972). Effect of cucumber mosaic virus on chlorophyll content, photosynthesis, respiration and carbohydrates of infected *Luffa aegyptiaca* Mill. *Journal of Phytopathology*, 75, 181–186.
- Manfre, A., Glenn, M., Nunez, A., Moreau, R., & Dardick, C. (2011). Light quantity and photosystem function mediate host susceptibility to turnip mosaic virus via a salicylic acid independent mechanism. *Molecular Plant-Microbe Interactions*, 24(3), 315–327.
- Martin, G. E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., et al. (2014). The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: Evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. *Annals in Botany*, 113, 1197–1210.
- Mazidah, M., Lau, W. H., Yusoff, K., Habibuddin, H., & Tan, Y. H. (2012). Ultrastructural features of *Catharanthus roseus* leaves infected with cucumber mosaic virus. *Pertanika Journal of Tropical Agricultural Science*, 35, 85–92.
- Mochizuki, T., Ogata, Y., Hirata, Y., & Ohki, S. T. (2014). Quantitative transcriptional changes associated with chlorosis severity in mosaic leaves of tobacco plants infected with cucumber mosaic virus. *Molecular Plant Pathology*, 15(3), 242–254.
- Morita, M. T., & Shimada, T. (2014). The plant endomembrane system—A complex network supporting plant development and physiology. *Plant and Cell Physiology*, 55(4), 667–671.
- Moore, M. J., Dhingra, A., Soltis, P. S., Shaw, R., Farmerie, W. G., Folta, K. M., et al. (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, 6, 17. Available from <https://doi.org/10.1186/1471-2229-6-17>.
- Moshe, A., Pfannstiel, J., Brotman, Y., Kolot, M., Sobol, I., Czosnek, H., et al. (2012). Stress responses to tomato yellow leaf curl virus (TYLCV) infection of resistant and susceptible tomato plants are different. *Metabolomics: Official Journal of the Metabolomic Society*, 51, 006. Available from <https://doi.org/10.4172/2153-0769.S1-006>.
- Mower, J. P. (2009). The PREP suite: Predictive RNA editors for plant mitochondrial genes, Chloroplast genes and user-defined alignments. *Nucleic Acids Research*, 37, W253–W259.
- Mur, L. A., Kenton, P., Lloyd, A. J., Ougham, H., & Prats, E. (2008). The hypersensitive response; the centenary is upon us but how much do we know? *Journal of Experimental Botany*, 59, 501–520.
- Musetti, R., Bruni, L., & Favali, M. A. (2002). Cytological modifications in maize plants infected by barley yellow dwarf virus and maize dwarf mosaic virus. *Micron (Oxford, England: 1993)*, 33(7–8), 681–686.
- Muthukumar, V., Melcher, U., Pierce, M., Wiley, G. B., Roe, B. A., Palmer, M. W., et al. (2009). Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. *Journal of Virology Methods*, 141, 169–173.
- Mwaba, I., & Rey, M. E. C. (2017). Nitric oxide associated protein 1 is associated with chloroplast perturbation and disease symptoms in *Nicotiana benthamiana* infected with South African cassava mosaic virus. *Virus Research*, 238, 75–83.
- Nashima, K., Terakami, S., Nishitani, C., Kunihiya, M., Shoda, M., Takeuchi, M., et al. (2015). Complete chloroplast genome sequence of pineapple (*Ananas comosus*). *Tree Genetics & Genomes*, 11, 60–71.
- Nomura, H., Komori, T., Uemura, S., Kanda, Y., Shimotani, K., Nakai, K., et al. (2012). Chloroplast-mediated activation of plant immune signalling in *Arabidopsis*. *Nature Communications*, 3, 926.
- Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., et al. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, 322, 572–574.
- Oldenburg, D. J., & Bendich, A. J. (2015). DNA maintenance in plastids and mitochondria of plants. *Frontiers Plant Science*, 6, 883.
- Oldenburg, D. J., & Bendich, A. J. (2016). The linear plastid chromosomes of maize: Terminal sequences, structures, and implications for DNA replication. *Current Genetics*, 62, 431–442.
- Otulak, K., Chouda, M., Bujarski, J., & Garbaczewska, G. (2015). The evidence of tobacco rattle virus impact on host plant organelles ultrastructure. *Micron (Oxford, England: 1993)*, 70, 7–20.
- Palmer, J. D., Osorio, B., & Thompson, W. F. (1988). Evolutionary significance of inversions in legume chloroplast DNAs. *Current Genetics*, 14, 65–74.
- Palmer, J. D., & Thompson, W. F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell*, 29, 537–550.
- Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., et al. (2017). Next generation sequencing for detection and discovery of plant viruses and viroids: Comparison of two approaches. *Frontiers in Microbiology*, 8, 1998–2008.
- Perez-Bueno, M. L., Rahoutei, J., Sajjani, C., Garcia-Luque, I., & Baron, M. (2004). Proteomic analysis of the oxygen-evolving complex of photosystem II under biotec stress: Studies on *Nicotiana benthamiana* infected with tobamoviruses. *Proteomics*, 4(2), 418–425.

- Pineda, M., Sajjani, C., & Baron, M. (2010). Changes induced by the pepper mild mottle tobamovirus on the chloroplast proteome of *Nicotiana benthamiana*. *Photosynthesis Research*, *103*, 31–45.
- Prodhomme, D., Jakubiec, A., Tournier, V., Drugeon, G., & Jupin, I. (2003). Targeting of the Turnip Yellow Mosaic Virus 66K replication protein to the chloroplast envelope is mediated by the 140K protein. *Journal of Virology*, *77*(17), 9124–9135.
- Provan, J., Powell, W., & Hollingsworth, P. M. (2001). Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends in Ecology and Evolution*, *16*, 142–147.
- Qiao, Y., Li, H. F., Wong, S. M., & Fan, Z. F. (2009). Plastocyanin transit peptide interacts with potato virus X coat protein, while silencing of plastocyanin reduces coat protein accumulation in chloroplasts and symptom severity in host plants. *Molecular Plant-Microbe Interactions*, *22*(12), 1523–1534.
- Rahoutei, J., Garcia-Luque, I., & Baron, M. (2000). Inhibition of photosynthesis by viral infection: Effect on PSII structure and function. *Physiologia Plantarum*, *110*, 286–292.
- Reinero, A., & Beachy, R. N. (1986). Association of TMV coat protein with chloroplast membranes in virus-infected leaves. *Plant Molecular Biology*, *6*, 291–301.
- Reinero, A., & Beachy, R. N. (1989). Reduced photosystem II activity and accumulation of viral coat protein in chloroplasts of leaves infected with tobacco mosaic virus. *Plant Physiology*, *89*, 111–116.
- Revers, F., & García, J. A. (2015). Molecular biology of potyviruses. In K. Maramorosch, & T. C. Mettenleiter (Eds.), *Advances in virus research* (Vol. 92, pp. 101–199). Amsterdam, The Netherlands: Academic Press, Elsevier Inc.
- Roberts, P. L., & Wood, K. R. (1982). Effects of a severe (P6) and a mild (W) strain of cucumber mosaic virus on tobacco leaf chlorophyll, starch and cell ultrastructure. *Physiological Plant Pathology*, *21*, 31–37.
- Rodríguez, A., Angel, C. A., Lutz, L., Leisner, S. M., Nelson, R. S., & Schoelz, J. E. (2014). Association of the P6 protein of cauliflower mosaic virus with plasmodesmata and plasmodesmal proteins. *Plant Physiology*, *166*(3), 2492–2500.
- Rodríguez, M., Muñoz, N., Lenardon, S., & Lascano, R. (2012). The chlorotic symptom induced by sunflower chlorotic mottle virus is associated with changes in redox-related gene expression and metabolites. *Plant Science*, *196*, 107–116.
- Roossinck, M. J. (2011). The good viruses: Viral mutualistic symbioses. *National Reviews of Microbiology*, *9*, 99–108.
- Roossinck, M. J., Martin, D. P., & Roumagnac, P. (2015). Plant virus metagenomics: Advances in virus discovery. *Phytopathology*, *105*, 716–727.
- Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., et al. (2010). Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology*, *19*, 81–88.
- Rowland, D., Dorner, J., Sorensen, R., Beasley, J. P., & Todd, J. (2005). Tomato spotted wilt virus in peanut tissue types and physiological effects related to disease incidence and severity. *Plant Pathology*, *54*(4), 431–440.
- Ruhlman, T., Lee, S. B., Jansen, R. K., Hostetler, J. B., Tallon, L. J., Town, C. D., & Daniell, H. (2006). Complete plastid genome sequence of *Daucus carota*: Implications for biotechnology and phylogeny of angiosperms. *BMC Genomics*, *7*, 222.
- Sablok, G., Mudunuri, S. B., Patnana, S., Popova, M., Fares, M. A., & Porta, N. L. (2013). ChloroMitoSSRDB: Open source repository of perfect and imperfect repeats in organelle genomes for evolutionary genomics. *DNA Research*, *20*(2), 127–133.
- Saski, C., Lee, S. B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H. G., et al. (2005). Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant Molecular Biology*, *59*(2), 309–322.
- Saski, C., Lee, S. B., Fjellheim, S., Guda, C., Jansen, R. K., Luo, H., et al. (2007). Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theoretical and Applied Genetics*, *115*, 571–590.
- Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., & Tabata, S. (1999). Complete structure of the chloroplast genome of *Thaliana ssc.* *DNA Research*, *6*, 283–290.
- Saxena, S., Kaila, T., Chaduvula, P. K., Singh, A., Singh, N. K., & Gaikwad, K. (2019). Novel chloroplast microsatellite markers in pigeonpea (*Cajanus cajan* L. Millsp.) and their transferability to wild *Cajanus* species. *Australian Journal of Crop Science*, *13*(2), 185–191.
- Shalla, T. A. (1964). Assembly and aggregation of tobacco mosaic virus in tomato leaflets. *Journal of Cell Biology*, *21*, 253–264.
- Shalla, T. A., Petersen, L. J., & Giunchedi, L. (1975). Partial characterization of virus-like particles in chloroplasts of plants infected with the U5 strain of TMV. *Virology*, *66*(1), 94–105.
- Shimura, H., Pantaleo, V., Ishihara, T., Myojo, N., Inaba, J., Sueda, K., ... Masuta, C. (2011). A viral satellite RNA induces yellow symptoms on tobacco by targeting a gene involved in chlorophyll biosynthesis using the RNA silencing machinery. *PLoS Pathogens*, *7*, e1002021.
- Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., et al. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. *The EMBO Journal*, *5*, 2043–2049.
- Smith, N. A., Eamens, A. L., & Wang, M. B. (2011). Viral small interfering RNAs target host genes to mediate disease symptoms in plants. *PLoS Pathogens*, *7*, e1002022.
- Sun, X., Li, Y., Shi, M., Zhang, N., Wu, G., Li, T., et al. (2013). *In vitro* binding and bimolecular fluorescence complementation assays suggest an interaction between tomato mosaic virus coat protein and tobacco chloroplast ferredoxinI. *Archives of Virology*, *158*(12), 2611–2615.
- Tangphatsornruang, S., Sangsrakru, D., Chanprasert, J., Uthapaisanwong, P., Yoocha, T., Jomchai, N., et al. (2010). The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: Structural organization and phylogenetic relationships. *DNA Research*, *17*, 11–22.
- Thiel, T. (2003). MISA—Microsatellite identification tool. <http://pgrc.ipk-gatersleben.de/misa/>.
- Thomas, F., Massenet, O., Dorne, A. M., & Briat, J. M. R. (1988). Expression of the rpl23, rpl2, and rps19 genes in spinach chloroplasts. *Nucleic Acids Research*, *16*, 2461–2472.

- Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq – Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*, W6–W11.
- Toivola, D. M., Strnad, P., Habtezion, A., & Omary, M. B. (2010). Intermediate filaments take the heat as stress proteins. *Trends in Cell Biology*, *20*(2), 79–91.
- Torrance, L., Cowan, G. H., Gillespie, T., Ziegler, A., & Lacomme, C. (2006). Barley stripe mosaic virus-encoded proteins triple-gene block 2 and γ b localize to chloroplasts in virus-infected monocot and dicot plants, revealing hitherto-unknown roles in virus replication. *Journal of General Virology*, *87*(8), 2403–2411.
- Torres, M. A., Jones, J. D. G., & Dangl, J. L. (2006). Reactive oxygen species signaling in response to pathogens. *Plant Physiology*, *141*, 373–378.
- Vu, H. T., Tran, N., Nguyen, T. D., Vu, Q. L., Bui, M. H., Le, M. T., & Le, L. (2020). Complete chloroplast genome of *Paphiopedilum delenatii* and phylogenetic relationships among Orchidaceae. *Plants (Basel)*, *9*(1), 61.
- Wamonje, F. O., Michuki, G. N., Braidwood, L. A., Njuguna, J. N., Mutuku, J. M., Djikeng, A., et al. (2017). Viral metagenomics of aphids present in bean and maize plots on mixed-use farms in Kenya reveals the presence of three dicistroviruses including a novel Big Sioux River virus-like dicistrovirus. *Virology Journal*, *14*, 188.
- Wang, B., Hajano, J. U., Ren, Y., Lu, C., & Wang, X. (2015). iTRAQ-based quantitative proteomics analysis of rice leaves infected by rice stripe virus reveals several proteins involved in symptom formation. *Virology Journal*, *12*, 99.
- Wang, P., & Hussey, P. J. (2015). Interactions between plant endomembrane systems and the actin cytoskeleton. *Frontier in Plant Science*, *6*, 422.
- Wang, S., Cui, W., Wu, X., Yuan, Q., Zhao, J., Zheng, H., Lu, Y., Peng, J., Lin, L., Chen, J., et al. (2018). Suppression of nbe-miR166h-p5 attenuates leaf yellowing symptoms of potato virus X on *Nicotiana benthamiana* and reduces virus accumulation. *Molecular Plant Pathology*, *19*, 2384–2396.
- Wasternack, C. (2007). Jasmonates: An update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Annals of Botany*, *100*(4), 681–697.
- Wasternack, C., & Hause, B. (2013). Jasmonates: Biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in *Annals of Botany*. *Annals of Botany*, *111*(6), 1021–1058.
- Wildermuth, M. C., Dewdney, J., Wu, G., & Ausubel, F. M. (2001). Isochorismate synthase is required to synthesize salicylic acid for plant defence. *Nature*, *414*, 562–565.
- Williams, A. V., Boykin, L. M., Howell, K. A., Nevill, P. G., & Small, I. (2015). The complete sequence of the acacia ligulata chloroplast genome reveals a highly divergent clpP1 gene. *PLoS One*, *10*, e0125768.
- Wojciechowski, M. F., Sanderson, M. J., Steele, K. P., & Liston, A. (2000). Molecular phylogeny of the “Temperate Herbaceous Tribes” of papilionoid legumes: A supertree approach. *Advances in Legume Systematics*, *9*, 277–298.
- Wolf, S., Lucas, W. J., Deom, C. M., & Beachy, R. N. (1989). Movement protein of tobacco mosaic virus modifies plasmodesmatal size exclusion limit. *Science (New York, N.Y.)*, *246*, 377–379.
- Wolfe, K. H., Morden, C. W., & Palmer, J. D. (1992). Function and evolution of a minimal plastid genome from a non-photosynthetic parasitic plant. *Proceedings of the National Academy of Sciences of the United States of America*, *89*, 10648–10652.
- Worth, J., Liu, L., Wei, F. J., & Tomaru, N. (2019). The complete chloroplast genome of *Fagus crenata* (subgenus *Fagus*) and comparison with *F. engleriana* (subgenus *Engleriana*). *Peer Journal*, *7*, e7026.
- Wu, L., Wang, S., Chen, X., Wang, X., Wu, L., Zu, X., et al. (2013). Proteomic and phytohormone analysis of the response of maize (*Zea mays* L.) seedlings to sugarcane mosaic virus. *PLoS One*, *8*, e70295.
- Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics (Oxford, England)*, *20*(17), 3252–3255.
- Xiang, Y., Kakani, K., Reade, R., Hui, E., & Rochon, D. A. (2006). A 38-aminoacid sequence encompassing the arm domain of the cucumber necrosis virus coat protein functions as a chloroplast transit peptide in infected plants. *Journal of Virology*, *80*, 7952–7964.
- Xu, K., & Nagy, P. D. (2010). Dissecting virus-plant interactions through proteomics approaches. *Current Proteomics*, *7*, 316–327.
- Yan, L., Lai, X., Li, X., Wei, C., Tan, X., & Zhang, Y. (2015). Analyses of the complete genome and gene expression of chloroplast of sweet potato (*Ipomoea batata*). *PLoS One*, *10*, e124083.
- Yan, S. L., Lehrer, A. T., Hajirezaei, M. R., Springer, A., & Komor, E. (2008). Modulation of carbohydrate metabolism and chloroplast structure in sugarcane leaves which were infected by sugarcane yellow leaf virus (SCYLV). *Physiology and Molecular Plant Pathology*, *73*, 78–87.
- Zhang, C., Liu, Y., Sun, X., Qian, W., Zhang, D., & Qiu, B. (2008). Characterization of a specific interaction between IP-L, a tobacco protein localized in the thylakoid membranes, and tomato mosaic virus coat protein. *Biochemical and Biophysical Research Communications*, *374*(2), 253–257.
- Zhang, Q., & Liu, Y. (2003). Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant Cell Physiology*, *44*, 941–951.
- Zhao, J., Liu, Q., Zhang, H., Jia, Q., Hong, Y., & Liu, Y. (2013). The RubisCO small subunit is involved in Tobamovirus movement and Tm-22-mediated extreme resistance. *Plant Physiology*, *161*(1), 374–383.
- Zhao, J., Zhang, X., Hong, Y., & Liu, Y. (2016). Chloroplast in plant-virus interaction. *Frontiers in Microbiology*, *7*, 1565–1585.
- Zheng, X. Y., Spivey, N. W., Zeng, W., Liu, P. P., Fu, Z. Q., Klessig, D. F., et al. (2012). Coronatine promotes *Pseudomonas syringae* virulence in plants by activating a signaling cascade that inhibits salicylic acid accumulation. *Cell Host and Microbe*, *11*(6), 587–596.
- Zurbriggen, M. D., Carrillo, N., & Hajirezaei, M. R. (2010). ROS signaling in the hypersensitive response: When, where and what for? *Plant Signal Behavior*, *5*, 393–396.

Section III

Data mining, markers discovery

This page intentionally left blank

Deciphering soil microbiota using metagenomic approach for sustainable agriculture: an overview

Aiman Tanveer¹, Shruti Dwivedi¹, Supriya Gupta¹, Rajarshi Kumar Gaur² and Dinesh Yadav¹

¹Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India, ²Department of Biotechnology, Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India

27.1 Introduction

Soil is the most dynamic environment with huge plethora of known and unknown microbial species thriving in it. The complex biochemical mechanism occurring in the soil reflects several unknown functions which are very much important for sustenance of life (Sabale, Suryawanshi, & Krishnaraj, 2019). Soil erosion, extensive agricultural practices, climatic conditions, the everyday changing scenarios challenge the microbial diversity. To assess these microbial diversities and to understand their functions various approaches are used. These methods can extensively be classified under culture-dependent and culture-independent methods. The culture-dependent methods usually account various biochemical parameters to understand microbial diversity. Agricultural microbiomes are studied by targeting rhizospheric, endophytic, and phyllospheric microbiome. Microbial diversity in the soil have been projected to maintain the sustainability of agriculture production systems. The associated microbiomes are largely influenced by the environmental factors affecting the host plants such as the type and pH of soil, mineral content in soil, rainfall, salinity, temperatures. To know the diversity and distribution among different groups of microbes associated with different crops in the form of epiphytic, endophytic, and rhizospheric should be explored through culturable and unculturable techniques (Yadav, Kumar, Dhaliwal, Prasad, & Saxena, 2018). Since the microbial diversity plays integral role in fundamental metabolic processes in the soil, a basic understanding of diversity and function of soil biota is required in for preserving the integrity, function and long-term sustainability varied ecosystems (Sabale et al., 2019).

27.2 Sustainable agriculture

The land used for agriculture should be capable of maintaining their productivity indefinitely and should be useful to society for long term. This will be helpful in conserving the resources and would be useful to the environment and to the society. This has given way to the sustainable agricultural practices. The basic tenant of sustainable agriculture is preservation of environmental health, economic profitability, and maintenance of social equity.

Hence the production should not compromise with the human as well as natural resources. The human resources refer to the laborers, whose health and living conditions should be taken care of. The sustainable agriculture can be achieved with the integrated efforts of researchers, educationalists, policymakers, farmers, laborers, retailers, and the consumers. For achieving this objective, planned strategy should be adapted for internal cycling of nutrients and energy. This can also be obtained by minimizing the use of toxic chemicals and fertilizers

Soil is important natural resources that must be considered for long term sustainability. The highly diverse and dynamic soil microbial community is important for sustainable agriculture approach. The soil microbes are important in improving the soil health as well as the crop productivity. The diverse microbes associated with the crop in soil plays a major role in crop improvement. As it is already known that only 1% of the microbes are culturable under laboratory conditions and 99% still remain to be explored. Culture-independent techniques clubbed with the recent sequencing

technologies have helped in exploring the soil microbial communities. Although the soil is rich in the microbial population, but the area near the crop roots known as rhizosphere is specifically enriched with the microbes due to high nutrient content present there in. The microbial community present in the rhizosphere mainly comprises of bacteria, fungi, algae, nematodes protozoa and microarthropods (Raaijmakers, Paulitz, & Steinberg, 2009). Rhizosphere is one of the most extensively studied regions of the soil environment (Hiltner, 1904). This region comprises of huge microbial diversity and the microbial population in this region is around 10–100 times higher than bulk soil (Verma, Yadav, Khannam, & Mishra, 2019).

27.3 Soil microbiomes

Soil harbors varied microbial communities because it provides favorable condition as well as ecological niches for survival and supporting the metabolic activities of the thriving microbes. Soil is the main source of nutrient and raw material for plant growth. Sustainable agriculture relies on the strategies to improve or maintain the current rate of food production without damaging to the environmental and human health (Yadav et al., 2017).

Soil microbes are well recognized for maintain the balance between the diverse interacting factors responsible for the environmental equilibrium. Thereby standing as an integral element for a sustainable healthy food production. Microbes are primarily attracted to the rhizo deposit pools where they develop their microhabitat for survival (Hirsch, Miller, & Dennis, 2013). The in vivo influence of the soil microbial population to the environment has a greater impact than the artificially created microbial ecosystem for remediation purposes. The relative abundance of the microbial population has large impact on the ecosystem functioning, more, the alteration in relative abundance of organisms regulating the metabolic processes have direct effect on the rate of that very mechanism (Schimel & Schaeffer, 2012).

The rhizosphere is defined as the soil present in the vicinity of the roots (Hiltner, 1904). The microbial community structure in the rhizosphere is different from that found in the bulk soil. This may be due to either presence of the root exudates having high concentration of the nutrients or may be because of the increased microbial biomass which cause alteration in the environmental conditions of the rhizosphere. The rhizosphere is being incessantly influenced by plant roots through the rhizodeposition of exudates, mucilages, and sloughed cells (Bais, Weir, & Perry, 2006). Thus, plant roots can have an impact on the surrounding soil and the inhabiting microbial community. Mutually, the rhizospheric organisms can have influence on the plant by producing regulatory compounds. Thus, the rhizospheric microbiome acts as a highly evolved external functional milieu for plants (Badri, Weir, van der Lelie, &

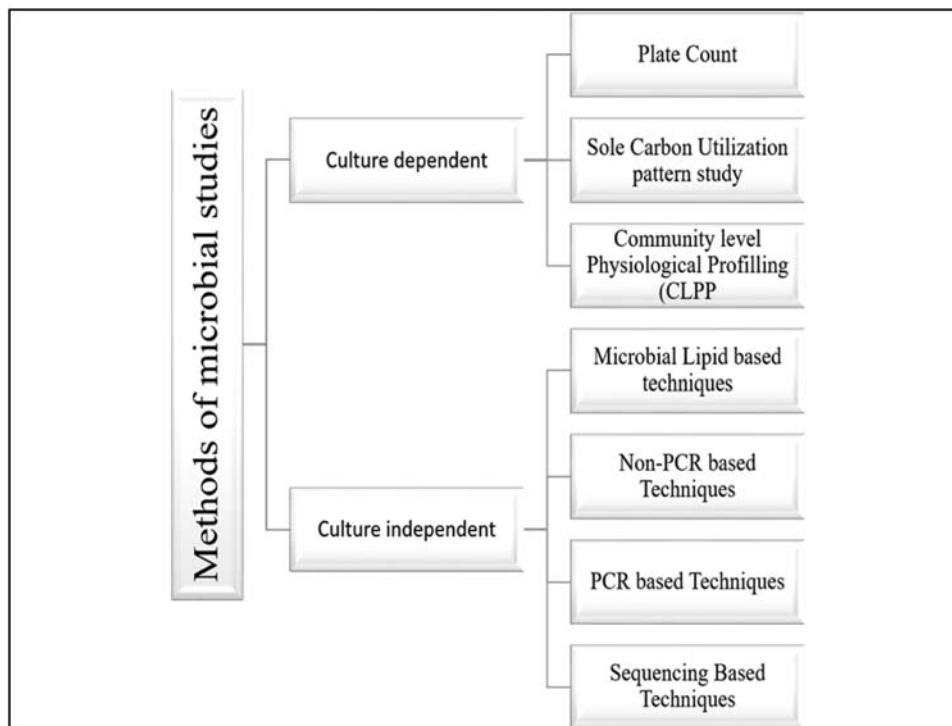


FIGURE 27.1 Common culture-dependent and culture-independent methods used for microbial analysis.

CULTURE- INDEPENDENT METHOD FOR MICROBIOME STUDIES

- Microbial Lipid based techniques**
 - PFLA(phospholipid fatty acid analysis)
 - FAME (Fatty Acid Methyl Ester)
- Non-PCR based Techniques**
 - DNA –reassociation
 - Guanine-Cytosine(G+C) content of DNA
 - RSGP (Reverse Sample Genome Probing)
- Sequencing BasedTechniques**
 - Clone library sequencing
 - Amplicon sequencing
 - Shotgun metagenome sequencing
- PCR based techniques**
 - Microsatellite region characterization
 - RAPD (Random Amplification of Polymorphic DNA)
 - RFLP (Restriction Fragment Length Polymorphism)
 - ARDRA (Amplified Ribosomal DNA Restriction Analysis)
 - T-RFLP (Terminal Restriction fragment length polymorphism)
 - RISA (Ribosomal intergenic spacer analysis)/ ARISA (Automated RISA)
 - TGGE/DGGE (Temperature gradient gel electrophoresis/ Denaturing gradient gel electrophoresis)
 - DNA Microarray
 - SSCP (Single stranded conformation polymorphism)
 - Real Time PCR (q PCR)

FIGURE 27.2 Culture-independent methods used for the microbiome analysis.

Vivanco, 2009; Bais et al., 2006). In this chapter we will be dealing with the metagenomics based analysis of the rhizospheric soil sample. The overview of culture-dependent and independent methods and techniques both conventional and recent are highlighted in Figs. 27.1 and 27.2.

27.4 Soil microbial diversity

Soil is the conglomerates of millions of fungi, billions of bacteria, and other macro organisms (Bardgett & Van Der Putten, 2014). Different type of soil have different diversity of microbes which is effect grow of plants. The soil microbe's effect differed depending on the soil type and the plant growth developmental stage (Schreiter, Sandmann, Smalla, & Grosch, 2014). Diverse microbial genera have been reported from rhizospheric soil of different host plants including *Arenimonas*, *Azotobacter*, *Bradyrhizobium*, *Burkholderia*, *Chitinophaga*, *Delftia*, *Dyella*, *Enterobacter*, *Erwinia*, *Flavobacterium*, *Gluconacetobacter*, *Klebsiella*, *Lysobacter*, *Massilia*, *Methylobacterium*, *Methylocystis*, *Ohtaekwangia*, *Paenibacillus*, *Pseudomonas*, *Sphingobium*, *Stenotrophomonas*, and *Variovorax* (Li et al., 2014), *Azospirillum*, *Bacillus*, *Flavobacterium*, *Micrococcus*, and *Staphylococcus* (Rameshkumar, Krishnan, Kandeepan, & Kayalvizhi, 2014), *Achromobacter*, *Acinetobacter*, *Agrobacterium*, *Alcaligenes*, *Arthrobacter*, *Duganella*, *Exiguobacterium*, *Kocuria*, *Lysinibacillus*, *Planococcus*, *Planomicrobium*, *Rhodobacter*, *Salmonella*, *Serratia*, *Sporosarcina*, and *Xanthomonas* (Majeed, Abbasi, Hameed, Imran, & Rahim, 2015), *Aspergillus* (Wang et al., 2018), *Penicillium* (Elias, Woyessa, & Muleta, 2016; Mittal, Singh, Nayyar, Kaur, & Tewari, 2008), *Talaromyces* (Zhang et al., 2018), and *Trichoderma* (Kapri & Tewari, 2010).

27.5 Analysis of the rhizosphere microbial community

To assess these microbial diversities and to understand their functions various approaches are used. These methods can extensively be classified under culture-dependent and culture-independent methods.

The culture-dependent methods usually account various biochemical parameters to understand microbial diversity. Abiotic and Biotic factors affect microbial diversity (Fakruddin & Mannan, 2013). Agricultural microbiome studies targets rhizospheric, endophytic, and phyllospheric microbiome. The method in general involves culturing microbe in specific media. For endophytic and phyllospheric surface sterilization is performed prior to microbes culturing. These cultured microbes can be assessed molecularly or biochemically or by both to achieve the objective of study (Yadav et al., 2018). This method of screening and isolation of microbe is easy, feasible, and result oriented. But the spatial heterogeneity of microbes, their dependence over abiotic and biotic factors, the risk of contamination in pure culture raises arrow towards some amendments to the methods in use. The culture-dependent method relies on the tools such as plate count, analysis of the carbon utilization pattern, and community level physiological profiling.

Culture-independent methods on contrary involves no culturing of the microbes. In this method, collective isolation of DNA takes place. The viable source of information regarding the soil microbes can be discovered through the biomolecules such as lipids, DNA, RNA, and proteins. The extraction procedures of the biomolecules from the soil is a

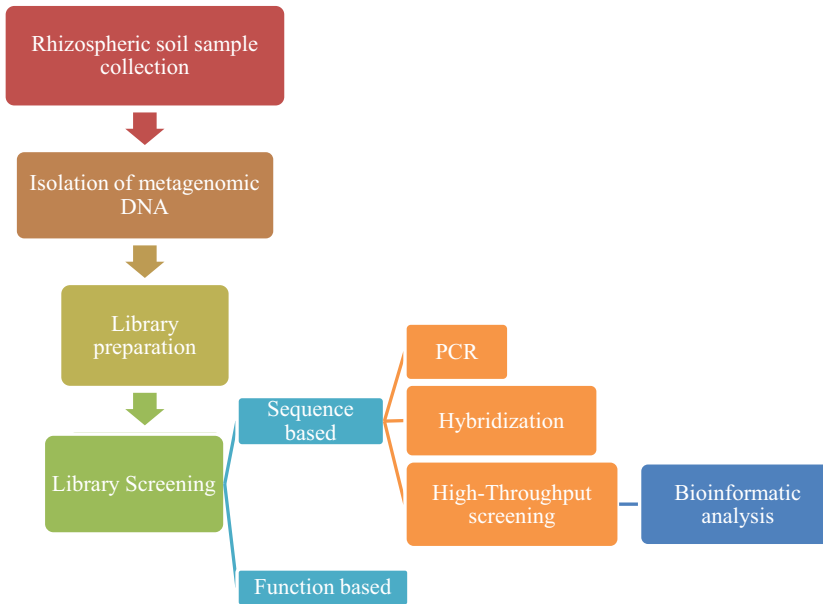


FIGURE 27.3 Workflow in the metagenomic analysis of rhizospheric soil sample.

challenging task due to variability and high concentration of the soil content, soil structure, and humic acids content. These parameters vary in its location and time. A new emerging area of collective isolation of DNA is called as metagenome. And this approach is coined as Metagenomics. This complete metagenomic DNA is processed for library cloning for functional and sequence based studies.

Culture-independent tools for the analysis of the microbial community includes lipid profiling techniques such as PFLA and FAME; PCR based techniques such as RAPD, RFLP, DGGE, qPCR, RISA to name a few; non PCR based techniques such as reverse sample probing technique, G + C content analysis. In the present chapter we will focus primarily on the metagenomics based approach for the analysis of the rhizospheric microbes.

27.6 Metagenomics in agriculture

Metagenomics has revolutionized the study of environmental microbiology and microbial ecology. The microbial diversity and function can be extensively studied ecological niches. It is a culture-independent method of assessing the largely untapped genetic reservoir microbial communities present in the soil environment. Metagenomics is performed through high-throughput sequencing technology to infer the taxonomic and functional traits of biological communities present in the environmental samples (Handelsman et al., 1998). It is performed without isolation of the microorganisms from the sample site. It directly detects and quantifies DNA. With the help of recent high through put sequencing platforms, the data can be quickly and accurately obtained. The obtained microbial data has been widely used in the analysis of soil microorganisms (Yuan et al., 2020).

The microbial diversity of soils serves as a promising source for exploring a wide range of industrial, agricultural, and environmental niches. This helps exploring the soil microbial communities extensively that can be useful in several aspects. These are in relevance to soil habitat, abundance, synergistic or antagonistic interactions, microbial diversity, their functional and phylogenetical interconnections, dynamic stability, and sensitivity of microbial communities (Simonet, Nesme, Achouk, & Agathos, 2016). Metagenomics is an integrated branch empowered with genomics, bioinformatics, and systems biology. It provides comprehensive overview of the microbial world. Hence genomics aspect has given way to the detailed analysis of the microbial communities (Sabale et al., 2019).

Meta science takes in account the aggregate genes of microbes, secondly it takes in account the computational biology tools to expedite the functional and sequential attribute. Metagenomics is sectioned into two major approaches, sequence based and function-based which target different features of the local microbial community within a determined environment.

The sequence-based approach is performed either by sequencing the clones of the library or random fragments obtained by metagenomic mining. 16S rDNA-based sequencing analysis of the clones gives an overview of the phylogenetic diversity of the metagenome (Soni, Shaluja, & Goel, 2010). The functional metagenomic approach leads to the

identification of the genes that code for a function of interest. This is done by activity-based screenings of the clones of the expression libraries (Guazzaroni et al., 2018). Metagenomics offers a new way of examining the microbial world that has given way to modern microbiology. It has potential to revolutionize understanding of the entire microbial living world and their functions in different ecosystem.

27.6.1 Metagenomics based techniques for rhizosphere analysis

The basic workflow for metagenomic analysis of the rhizospheric soil sample is shown in Fig. 27.3.

27.6.1.1 Sample collection and isolation of metagenomic DNA

For analysis of the rhizosphere, the bulk soil sample is separated and then the rhizospheric soil sample is collected from the roots of the crop plants. This is then subjected to DNA isolation. Isolation of good quality metagenomic DNA from soil is a prerequisite for a successful metagenomic study. Soil is a heterogeneous mixture comprising of high concentration of several contaminants like humic acid and phenolic compounds that inhibit the downstream processing of the DNA (Nair, Vincent, & Bhat, 2014). These contaminants coprecipitate with metagenomic DNA and make it unfit for further molecular based methods (Amorim et al., 2008). Several research groups have attempted to obtain high quality of the meta-DNA that is suitable for PCR and library preparation (Singh, Devi, Verma, & Rasool, 2014; Tanveer, Yadav, & Yadav, 2016; Volossiuk, Robb, & Nazar, 1995; Zhou, Bruns, & Tiedje, 1996).

27.6.1.2 Library preparation

Once a good quality DNA is obtained, the next step is the preparation of metagenomic library. The library preparation includes the cloning of the DNA fragments into specific vectors which are transformed into the host cell. These clones are then subsequently screened by either sequence or function-based approaches as described earlier. The vector is selected depending upon the size of DNA to be cloned.

Bacterial Artificial Chromosome (BAC)	100–200 Kb,
Cosmids	25–35 Kb
Fosmids	25–40 Kb
Yeast artificial chromosome (YAC)	over 40 Kb

Depending upon the strategy for the metagenomic DNA analysis, the size of the library is decided. Large size inserts are suitable for screening functional genes of high molecular weight. Host selection is also crucial for library preparation (Gabor, Alkema, & Janssen, 2004). *E. coli* is the most commonly used host for library preparation. It is easy to handle due to small genome size and high efficiency of transformation (Steele, Jaeger, Daniel, & Streit, 2009). But *E. coli* cannot express most of the genes present in the metagenome due to lack of expression machinery of large number of genes (Craig, Chang, Kim, Obiajulu, & Brady, 2010). To evade this, researchers have come up with alternative hosts such as *Bacillus*, *Pseudomonas*, and *Streptomyces* (Aakvik et al., 2009; Lorenz & Eck, 2005).

27.6.1.3 Library screening

The library thus obtained is then screened by either sequence and function-based approach. Each approach along the techniques associated with it will be described in this section.

27.6.1.3.1 Sequence-based screening

It involves sequencing of the clones of the library. This may be useful for diversity analysis, investigation of the genes present in the metagenome or for deciphering the phylogenetic ancestry of the microbes present therein.

The common techniques used for the above mentioned analysis are PCR based, probe based or high throughput sequencing based procedures.

1. PCR based screening

The library may be screened for specific enzymes and genes responsible for resistance to antibiotics. Specific primers are designed that amplify the gene of interest in the metagenome (Handelsman, 2004). The microbial community profiling may be performed with help of 16s rRNA-based amplification. Certain genes such as rRNA, recA, radA, nif, and phenol hydroxylase are used to decipher the phylogenetic relationship in the microbial population (Suenaga et al., 2009a).

2. Hybridization based screening

It allows simultaneous screening of large number of genes in the metagenomic DNA sample with the help of specific probes. The probes are designed to bind specific genes. Several labeled probes are also available that assist in the detection of the microbial wealth in the metagenome. Several microarray chips have been designed exclusively for the metagenome based analysis of the microbial community.

GeoChip is the example of fixed oligo microarray. The first version of the chip contained probes obtained by direct amplification of environmental DNA using primers for *nirS*, *nirK*, and *amoA* genes (Wu et al., 2001). Since then the number of probes in the chip has increased there by providing better analysis of the metagenomic community. The latest version (GeoChip 4.0) has 83,992 probes with 152,414 target genes which are divided into 410 categories. It covers the functional genes from fungi, archaea, bacteria, and viruses (Van Nostrand, He, & Zhou, 2012). GeoChip version 4.0 has been used in various analyses of metagenomic samples from the Amazon rainforest (Paula et al., 2014), and samples from effluent treatment plants (Wang et al., 2014). Virochip has been developed for the Screening of the viruses. Its probe is derived from the conserved sequences of several viral families (Wang et al., 2002). Similarly Human Gut Chip (HuGChip) and Human Intestinal Tract chip (HITChip) have been designed for the analysis of the gut microflora (Rajilić-Stojanović, Smidt, & De Vos, 2007; Tottey et al., 2013). Chip for antibiotic-resistant gene screening has been designed that contains 8746 probes for the 9 major groups of resistant genes. This chip has been used for characterization of the microflora and its association with the population of different age groups.

3. High-throughput sequencing-based screening

This allows large scale sequencing of the metagenomic DNA. This helps in functional profiling and community structure analysis for the sample collected from a particular site. The sequencing can be performed by either directly sequencing the metagenomic DNA or the clone collection obtained in the form of library. This helps in the discovery of new genes from the environmental samples. But it does not give the functional perspective of the genome. Hence function-based screening is performed to get functional insight of the metagenome.

The next-generation sequencing approach has revolutionized the analysis of the DNA obtained directly from the environmental samples. Several sequencing platforms with recent technologies have been introduced that facilitate the process. Advanced sequencing platforms used for sequence based screening of the metagenomic DNA library are listed in Table 27.1.

Roche 454 genome sequencer is one of the earliest tools used for sequence based analysis of the metagenomic library. It relies on sequence by synthesis methodology and the PPI released by the incorporation of the nucleotide is

TABLE 27.1 Advanced sequencing platforms used for sequence based screening of the metagenomic DNA library.

S. No	Technique	Read length	Properties
1.	Roche 454 genome sequence	up to 1000–1200 bp	highest cost per base and the lowest output
2.	Illumina sequencing (Solexa genome analyzer)	Up to 300bp, sequence 1 GB data in single run	low cost per base and high yield, multiplexing
3.	Applied biosystems (AB) SOLiD sequencer	85 bp	whole genome sequencing, targeted sequencing, transcriptome, and epigenome analysis
4.	Ion torrent sequencing	~200bp	sequencing quality is high and stable
5.	Helicos biosciences (HeliScope)	Upto 55bp /run Median read length of 35nt.	Sequencing without amplification, small read length (24–70 bases) and low data output (20 GB)
6.	PacBio technology/ SMRT sequencer	1500bp-20kb	fast sample preparation, no need for PCR amplification during the preparation, longer-read length than any other next-generation sequencing platform
7.	Oxford Nanopore technology	500bp-2.3 mb	read long sequences at low-cost in real time

detected. The sequencing takes place along with the amplification with the barcoding primers. Illumina sequencing is the most widely used sequencing platform for metagenomic study. Each incorporated nucleotide generates a luminescent signal which is recorded by the optical sensors.

In another sequencing strategy, applied biosystems (ABI) SOLiD sequencer. The sequence data is generated through signal detection by ligation with interrogation probes. Another sequencing platform is the Ion torrents sequencing, in which chemical sensors is used that detect the change in hydrogen ion concentration on incorporation of the nucleotide. The sequencing speed is high and the cost is comparatively low as compared to previously mentioned sequencing techniques (Liu et al., 2012). HeliScope platform allows sequencing without PCR amplification from single stranded DNA or RNA samples (Harris, Buzby, & Babcock, 2008; Zhang, Chiodini, Badr, & Zhang, 2011). The fluorescently labeled bases are attached at a time for sequencing along with the DNA polymerase. This generates a fluorescent signal which is detected by the CCD camera. Single-molecule real-time or PacBio technology is a real time sequencing technology which does not need PCR amplification. Zero-mode waveguide is used for observing the DNA synthesis in real time. It differs from other sequencing platforms that it uses different colored fluorescent labels for each nucleotide and the label is present at the terminal phosphate group of the nucleotide. Hence the signal is release by base incorporation (Flusberg et al., 2010). The washing step between nucleotide is not required needed thereby the quality of sequencing is improved and it can read longer bases as compared to other next-generation sequencing platforms (Zhou et al., 2010). Oxford Nanopore Technology devised by Nanopore sequencing is carried out by passing the DNA sequence through 1 nm diameter hole (nanopore) where electric current is applied. The current generated varies for each nucleotide and the signal is generated in real time (Hart et al., 2010). The MinION can read long sequences at low-cost in real time (Hayden, 2012). It is portable, pocket size and so the sample can be directly sequences at the sampling site itself (Jain et al., 2015).

The result generated by different sequencing strategies is subjected to bioinformatics analysis. The sequencing data can be used either to study the microbial diversity or the functional gene in the genome. 16S rRNA sequence strategy is used for understanding the microbial diversity at molecular level. QIIME, MOTHUR, DADA2, UPARSE are commonly used bioinformatics tools for analysis of 16S rRNA (Niu et al., 2018).

Few recent examples of analysis of the crop rhizospheric soil by sequence based metagenomics approach has been reported. Cotton rhizosphere soils from Alwar district of Rajasthan was subjected to library preparation and subsequent analysis (Singh, Johri, & Dua, 2020). Of the 185,231 assembled scaffolds 229,401 genes were predicted. These genes had an average gene length of 351 bp. About 75% of the reads could be identified taxonomically. Among the 75% identified reads, *Bacteria* (70%) dominated the microbial diversity, while *Eukaryota* (2%) and *Archaea* (3%) formed very small parts. Microbial communities analysis and its interaction with the plant crop in the rhizosphere of kodo millet (*Paspalum scrobiculatum*) was performed though metagenomics (Prabha et al., 2019). Microbial analysis has revealed that Actinobacteria and Proteobacteria were most abundant in the soil. The functional analysis has shown that the proteins present in the rhizospheric microbe encoded genes are involves in several mechanisms. The actinobacteria was shown to be mainly associated with the genes responsible for survival in the stressed and nutrient deprived conditions. Thus the rhizosphere associated microbes are responsible for survival under harsh conditions. Next-generation sequencing has been performed to analyze the rhizosphere microbial community of maize plants and comparative study was performed from plants grown in organic and inorganic manure (Enebe & Babalola, 2020). Proteobacteria and Bacteroidetes were present in both the sets but with different proportions. They have concluded that microbial community structure, relative abundance of each microbe and dynamics may help in the management of soil microbial community for improved agroecosystem.

27.6.1.3.2 Screening-based on function

Function-based screening provides the specific enzymatic activity of the screened clones of the library. It can be performed by four basic strategies as described by Felczykowska, Bloch, Nejman-Falencyk, & Baranska, 2012.

1. Performing enzymatic assays for specific enzymes
2. Induced gene expression
3. Direct detection of enzyme activity through fluorescent catabolic products
4. Heterologous complementation
5. Substrate-induced gene expression screening (SIGEX) (Ko, Han, Cheong, Choi, & Song, 2013)

The main challenge for function-based screening is that the functional gene expression is governed by several factors such as the incomplete genes cloning, incompatible expression factors, differences in codon usage, difference in protein folding pattern and lack of effective means to screen large number of clones, host incompatibly (Felczykowska et al., 2012; Schoenfeld et al., 2010).

27.7 Metagenomics for sustainable agriculture

Soil microorganism play an important role in nutrient cycling and protecting the plant from harmful abiotic and biotic stresses (Ahmad et al., 2012, 2013; Hashem, Abd_Allah, Alqarawi, Radhakrishnan, & Kumar, 2017; Kumar et al., 2010, 2013). Microbes play integral role in plant growth and functioning of ecosystems. For instance, Mycorrhizae, lives in symbiotic partnerships between plant roots and specific soil fungi which is grow in close association with the plant roots and even they can grow partially within the plant’s cells. Most plants are dependent on these mycorrhizal associations to obtain nutrients and to defend themselves against disease-causing microbes. The complete overview of the metagenomic analysis of the rhizospheric soil sample and its diverse applications in sustainable agriculture in present in Fig. 27.4. Application of the microbe is sustainable agriculture in listed in Table 27.2.

The soil matrix is the most biodiverse ecosystem on earth as it harbors large number of microbes that interact with plants (Vogel et al., 2009). The soil microbial diversity is directly linked to plant health. For instance, the soil microbes may suppress plant diseases by preventing the plant pathogen from infecting plant tissues (Mendes et al., 2011; Weller, Raaijmakers, Gardener, & Thomashow, 2002). It also restricts survival of exogenous organisms thereby maintaining the microbial integrity in and environmental niche (Elias, Woyessa, & Muleta, 2016; van Elsas et al., 2012) It is also now established that large number of the microbes present in the plant soil are not the cause of disease occurrence (Mendes, Garbeva, & Raaijmakers, 2013). Plant associated microbes serve several important functions namely (1) phosphorus solubilization and nitrogen fixation; (2) nutrient uptake; (3) Promote plant protection from biotic and abiotic stress (Halmann, Quadt-Hallmann, Mahaffee, & Kloepper, 1997; Mendes et al., 2011; Mendes et al., 2013; Quecine et al., 2014).

The conventional method of microbial study allows only cultivation of only 1% to 5% of the microbes by standard cultivation (Amann, Ludwig, & Schleifer, 1995). Exploration of the plant microbiomes interaction may reveal the whole set of microorganisms interacting with plants. This may also provide insight into numerous other functions that microbes exert when interacting with the host plant. Some authors have also hypothesized that a coevolution process

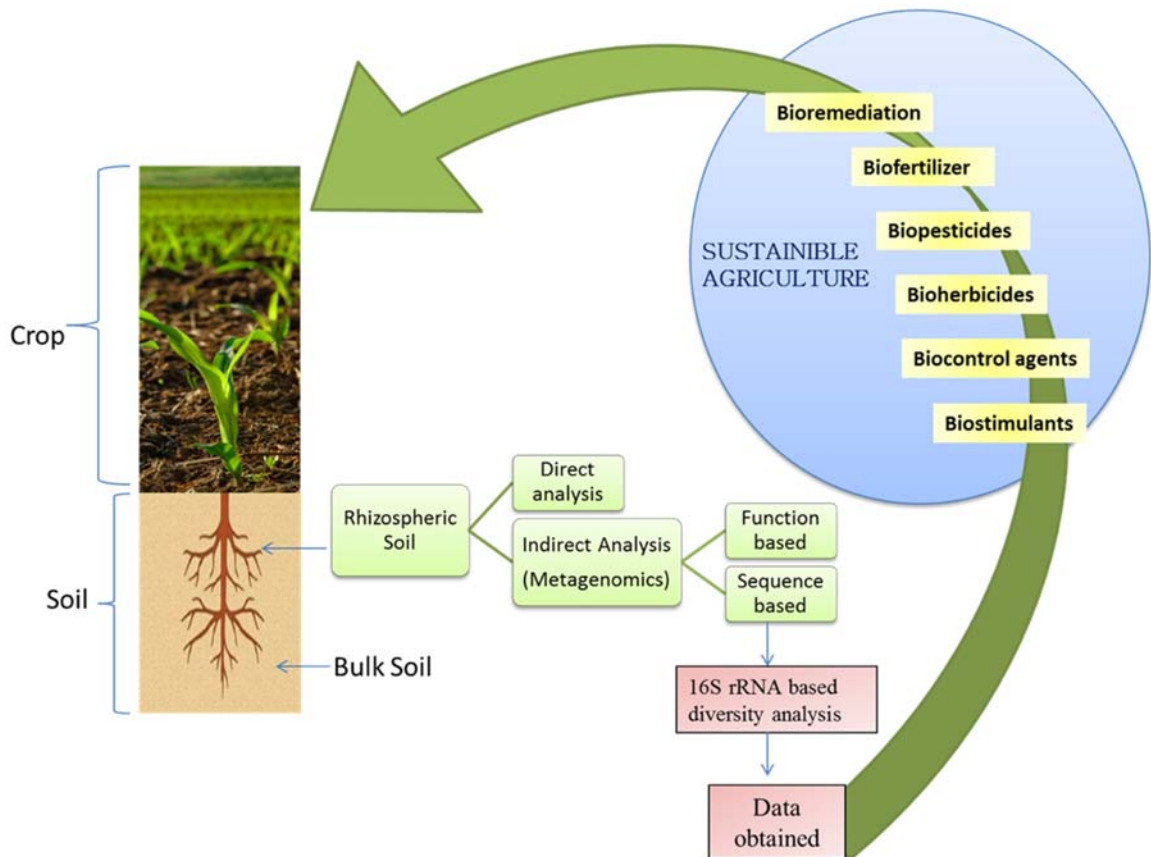


FIGURE 27.4 Overview of the metagenomic analysis of the rhizospheric soil sample and its application in sustainable agriculture.

TABLE 27.2 Application of soil rhizosphere microbiome in sustainable agriculture.

Uses	Sl. No.	Plants	Soil microbiomes	Function	References
BIO-FERTILIZER	1	Alder (<i>Alnus glutinosa</i> [L.] Gaertn.).	<i>Bacillus licheniformis</i> and <i>Bacillus pumilus</i>	produce gibberellins that helps in the promotion of plant growth	Guti_erez-Mañero et al. (2001)
	2	Peanut	<i>Pseudomonas fluorescens</i> , <i>Pseudomonas</i> spp.	Produce siderophore, ammonia, and indole acetic acid (IAA), solubilize the tri-calcium phosphate, ACC deaminase that considerably increases the root length of the seedlings of peanut.	Dey, Pal, Bhatt, and Chauhan (2004)
	3	barley, chickpea, maize, and pea	<i>Acinetobacter rhizosphaerae</i>	phosphate-solubilizing in plant, producing auxin and siderophore, ammonia, showing the activity of ACC deaminase activity	Gulati, Vyas, Rahi, and Kasana (2009)
	4	Wheat	<i>Bacillus thuringiensis</i> , <i>Enterobacter asburiae</i> , and <i>Serratia marcescens</i>	producing IAA, HCN, ammonia, and solubilizing phosphorus	Selvakumar, Kundu, Gupta, Shouche, and Gupta (2008)
	5	Tomato	<i>Pseudomonas aeruginosa</i>	Production of IAA, siderophore, solubilization of inorganic phosphate with the activity of with chitinase, urease, and b-1–3-glucanase.	Kumar, Pandey, and Maheshwari, (2009)
	6	Rice	<i>Agromyces</i> , <i>Bacillus</i> , <i>Microbacterium</i> , <i>Methylophaga</i> , and <i>Paenibacillus</i>	production of IAA, siderophore, ammonia and ACC deaminase activity	Bal, Das, Dangar, and Adhya (2013)
BIO-PESTICIDES	1	Potato, Pear, apple and other rosaceous plants	<i>Pseudomonas fluorescens</i> , <i>Erwinia herbicola</i>	Control the fire blight by suppressing the <i>Phytophthora infestans</i> and <i>Erwinia amylovora</i> .	Kumari, jha, kumar, and Rajanikant (2018)
	2	Bean	<i>B. subtilis</i>	Control the bean rust by resistance against <i>Uromyces</i> sp.	Kumari et al. (2018)
	3	Cruciferae	<i>S. griseoviridis</i>	Reducing the damping off of crucifer by suppressing the agent <i>Agrobacterium brasicaicola</i>	Kumari et al. (2018)
	4	Cotton	<i>P. fluorescens</i> <i>Rhizoctonia solani</i>	Prevent the growth of <i>Rhizoctonia solani</i> <i>P. ultimum</i> and reduces the damping off of cotton.	Kumari et al. (2018)
	5	Mushroom	<i>P. fluorescens</i>	Prevent the <i>Pythium ultimum</i> which cause the Brown blotch of Mushrooms	Kumari et al. (2018)
	6	Several crops (tomato, cotton, sugar beet, grapes etc.)	<i>A. radiobacter</i>	Control the bacterium <i>Agrobacterium tumefaciens</i> which causes crown gall.	Kumari et al. (2018)
	7	Cotton, chickpea, maize, tomato, groundnut etc.	<i>Bacillus thuringiensis</i>	It produced a toxin that specifically Kill the <i>Heliothis</i> and other Lepidoptera and Coleopteran.	Kumari et al. (2018)

(Continued)

TABLE 27.2 (Continued)

Uses	Sl. No.	Plants	Soil microbiomes	Function	References
	8	Citrus fruits plants	1. <i>Hirsutella thompsonii</i> 2. <i>Verticillium lecanii</i>	1. Controls the citrus rust mites. 2. control the Aphids, white, Lies	Kumari et al. (2018)
	9	Groundnut, chickpea	<i>Trichoderma viride</i>	Prevent the growth of fungus <i>Macrophomina phaseolina</i> which causes damping off, seedling blight, collar rot etc.	Kumari et al. (2018)
	10	Sisam	<i>T. viride</i>	Prevent the growth of <i>F. solani</i> that causes wilt.	Kumari et al. (2018)
	11	Rice, Cotton, Cabbage	<i>Nucleopolyhedrosis virus</i>	Suppressed the Rice borer, cotton leaf worm, and cabbage looper. It generally commercially use in USA.	Kumari et al. (2018)
12	Potato, rice	<i>Granulosis viruses (GV)</i>	Prevent Codling moth, tuber worm rice borer	Kumari et al. (2018)	
BIO STIMULANTS	1	Broccoli	<i>Brevibacillus reuszeri/Rhizobium rubi</i>	Promoting in the growth of the root system, an increased yield and an enhanced macro- and micronutrients uptake.	Yildirim, Karlidag, Turan, Dursun, and Goktepe (2011)
	2	Lettuce	<i>R. leguminosarum</i> bv. <i>phaseoli</i> strain P31	Promoting in the growth of the root system, an increased yield and an enhanced macro- and micronutrients uptake	Chabot, Antoun, and Cescas (1996)
	3	Fruit crop (apricot, apple cherry, banana)	<i>Bacillus</i> sp.	increase the production, the weight and the quality parameters in the aforementioned fruits	Esitken, Pirlak, Turan, and Sahin (2006), Kavino, Harish, Kumar, Saravanakumar, and Samiyappan (2010), Ryu et al. (2011)
	4	Apple	<i>Pseudomonas</i> spp.	Increases the yield	Aslantas, Çakmakçı, and Şahin (2007)
BIO HERBICIDES	1	Tomato	<i>Pseudomonas fluorescense</i>	Promote the growth of plants	Gamalero et al. (2005)
	2	Wheat	<i>Azotobacter</i> sp.	Stimulate the plant growth by nitrogen fixation.	Vessey (2003)
	3	Grass	<i>Pseudomonas fluorescens</i>	Help in the production antifungal substance that root inhibition in weed	Kremer (2019).
	4	Crops	<i>Phytophthora citrophora</i>	Help to control the growth of Milk weed	Kumari et al. (2018).
	5	Crops	<i>Colletrotrichum gloeosporioides</i>	Inhibit the <i>Aeschynomene virginica</i> growth.	Kumari et al. (2018).
	6	Crops	<i>Malameba locustae</i>	Control the Grass hoper, Lepidoptera in crops.	Kumari et al. (2018).
BIOCONTROL AGENTS	1	eggplant	<i>Pseudomonads</i>	Eggplant wilt caused by <i>Ralstonia solanacearum</i> was reduce.	Ramesh, Joshi, and Ghanekar (2008)

occurs between plants and its associated microbiome causing resilient genomic interdependency, leading to the “meta-organism” concept (Bosch & McFall-Ngai, 2011).

Rhizosphere associated microbes of agriculturally important crops can augment plant growth and improve plant nutrition through biological N₂ fixation and other mechanisms (Yadav et al., 2017). Microbes are associated with important functions such as increasing crop yields, removing of contaminants, inhibiting pathogens, and production of new substances (Quadt-Hallmann, Kloepper, & Benhamou, 1997). The growth stimulation by microbes can be a consequence of biological N₂-fixation (de Bruijn, Stoltzfus, So, Malarvithi, & Ladha, 1997; Iniguez, Dong, & Triplett, 2004; Pankievicz et al., 2015; Suman et al., 2001; Taulé et al., 2012); production of phytohormones, such as indole-3-acetic acid (IAA) and cytokinins (Lin & Xu, 2013; Rashid, Charles, & Glick, 2012;); biocontrol of phytopathogens through the production of antifungal or antibacterial agents (Errakhi, Bouteau, Barakate, & Lebrihi, 2016; Raaijmakers, Vlami, & De Souza, 2002); siderophores production (Ellis, 2017; Leong, 1986); nutrient competition (Bach, dos Santos Seger, de Carvalho Fernandes, Lisboa, & Passaglia, 2016); and induction of acquired host resistance (Van Loon, Bakker, & Pieterse, 1998), enhancing the bioavailability of minerals (Haas & Défago, 2005). The microbial plant growth promoters have been useful alternative to the conventional agricultural technologies (Kaur, Sharma, Chhabra, Chand, & Mangat, 2017) Yadav et al., 2017). They influence the plant growth directly or indirectly. The PGP microbes promote direct growth of plant by facilitating the uptake of certain nutrients from the environment. Indirectly the PGP microbes decrease or prevent the damaging effects of phytopathogenic organisms. The interactions of plants and microbes in the rhizosphere influence soil fertility and plant health.

The rhizospheric microbiomes may play promising role in stimulating the plant growth under the normal and abiotic stress conditions and has prospective to replace the harmful agrochemicals in the future (Yadav, 2020; Kour et al., 2020). Many plant growth-promoting microbes have been reported from stressed environmental conditions. *Pseudomonas* sp. and *Enterobacter* sp. (Sandhya et al., 2010), *Bacillus* sp., and *Paenibacillus* sp. (Vardharajula, Zulfikar Ali, Grover, Reddy, & Bandi, 2011) have been isolated from drought stress conditions. *Enterobacter* sp. (Sarkar et al., 2018), *Arthrobacter* sp., *Bacillus* sp., and *Pseudomonas* sp. (Upadhyay, Singh, & Saikia, 2009) from salinity stress. *Bacillus* sp., *Methylobacterium* sp., and *Pseudomonas* sp. (Verma et al., 2015) have been obtained from low-temperature stress conditions. The rhizospheric microbiomes assist the plant in uptake of the essential nutrients from the soil which cannot be easily taken up by the roots of the plant. These nutrients include phosphorus, potassium, zinc, and manganese.

Organic farming is one of the examples of the use of natural microflora. These soil microbes constitute of all kinds of useful bacteria and fungi including the arbuscular mycorrhiza fungi called plant-growth-promoting rhizobacteria (PGPR). A key advantage of beneficial microorganisms is to assimilate phosphorus for their own requirement, which in turn is available as its soluble form in sufficient quantities in soil. The microorganisms *Bacillus*, *Pseudomonas*, *Micrococcus*, *Flavobacterium*, *Fusarium*, *Sclerotium*, *Aspergillus*, and *Penicillium* have been reported which are active in the solubilization process. Many tools of modern science have been broadly connected for crop improvement under stress, of which PGPRs role as bioprotectants has turned out to be significantly useful (Yang, Kloepper, & Ryu, 2009).

Microbes assist in fixing atmospheric nitrogen for the plants and producing plant growth regulators including auxins, cytokinins, and gibberellins. These microbes also protect the plants from phytopathogens through siderophores which are low-molecular weight iron-chelating compounds. The plant growth-promoting microbes also help the plants to overcome abiotic and biotic stress conditions.

In a recent study, nematicidal microbes have been studied that act on the Root-knot (Niu, Paulson, Zheng, & Kolter, 2017) nematode *Meloidogyne incognita* which adversely affect the crop productivity (Zhao et al., 2021). The microbes associated with soil were deciphered through sequence based metagenomics by Illumina HiSeq platform. These microbes produced enzymes, proteases, chitinase, and lipases which act on the nematode causing its death.

27.8 Concluding remarks

With the diverse application of the rhizospheric microbes, the microbes serve as promising tool in sustainable agriculture. These rhizospheric microbes are highly beneficial for plant productivity due to synergistic interaction between the plants and the corresponding microbes. Many research groups are engaged in the analysis of microbial consortia present in the rhizosphere of specific crops. The current agriculture practices should be focused on adopting sustainable practices. It should be focused towards conserving the environment, reducing global warming and meeting the global food demands. The Food and Agriculture Organization of the United Nations estimated that the global food production needs to be increased by about 70% by 2050, so as to feed the projected world population of about 10 billion. The current

challenge is exploiting the sustainable path to produce more food in an environment friendly. That can cater the basic need of all on the planet.

References

- Aakvik, T., Degnes, K. F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L., . . . Valla, S. (2009). A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiology Letters*, *296*, 149–158.
- Ahmad, P., Hameed, A., Abd-Allah, E. F., Sheikh, S. A., Wani, M. R., Rasool, S., . . . Kumar, A. (2013). Biochemical and molecular approaches for drought tolerance in plants. In P. Ahmad, & M. R. Wani (Eds.), *Physiological mechanisms and adaptation strategies in plants under changing environment* (pp. 1–29). New York: Springer.
- Ahmad, P., Kumar, A., Ashraf, M., & Akram, N. A. (2012). Salt-induced changes in photosynthetic activity and oxidative defense system of three cultivars of mustard (*Brassica juncea* L.). *African Journal of Biotechnology*, *11*(11), 2694–2703.
- Amann, R. L., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiology Reviews*, *59*, 143–169.
- Amorim, J. H., Macena, T. N., Lacerda, G. V., Jr, Rezende, R. P., Dias, J. C., Brendel, M., & Cascardo, J. C. (2008). An improved extraction protocol for metagenomic DNA from a soil of the Brazilian Atlantic Rainforest. *Genetics and Molecular Research: GMR*, *7*(4), 1226–1232. Available from <https://doi.org/10.4238/vol7-4gmr509>, PMID: Available from 19065757.
- Aslantas, R., Çakmakçı, R., & Şahin, F. (2007). Effect of plant growth promoting rhizobacteria on young apple tree growth and fruit yield under orchard conditions. *Scientia Horticulturae*, *111*, 371–377.
- Bach, E., dos Santos Seger, G. D., de Carvalho Fernandes, G., Lisboa, B. B., & Passaglia, L. M. P. (2016). Evaluation of biological control and rhizosphere competence of plant growth promoting bacteria. *Applied Soil Ecology*, *99*, 141–149.
- Badri, D. V., Weir, T. L., van der Lelie, D., & Vivanco, J. M. (2009). Rhizosphere chemical dialogues: plant-microbe interactions. *Current Opinion in Biotechnology*, *20*, 642–650.
- Bais, H. P., Weir, T. L., Perry, L. G., et al. (2006). The role of root exudates in rhizosphere interactions with plants and other organisms. *Annual Review of Plant Biology*. Available from <https://doi.org/10.1146/annurev.arplant.57.032905.105159>.
- Bal, H. B., Das, S., Dangar, T. K., & Adhya, T. K. (2013). ACC deaminase and IAA producing growth promoting bacteria from the rhizosphere soil of tropical rice plants. *Journal of Basic Microbiology*, *53*, 972–984.
- Bardgett, R. D., & Van Der Putten, W. H. (2014). Belowground biodiversity and ecosystem functioning. *Nature*, *515*, 505–511. Available from <https://doi.org/10.1038/nature13855>.
- Bosch, T. C., & McFall-Ngai, M. J. (2011). Metaorganisms as the new frontier. *Zoology Review*, *114*, 185–190.
- Chabot, R., Antoun, H., & Cescas, M. P. (1996). Growth promotion of maize and lettuce by phosphate-solubilizing *Rhizobium leguminosarum* biovar. phaseoli. *Plant and Soil*, *184*, 311–321.
- Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C., & Brady, S. F. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Applied and Environmental Microbiology*, *76*, 1633–1641.
- de Bruijn, F., Stoltzfus, J., So, R., Malarvithi, P., & Ladha, J. (1997). *Isolation of endophytic bacteria from rice and assessment of their potential for supplying rice with biologically fixed nitrogen. Opportunities for Biological Nitrogen Fixation in Rice and Other Non-Legumes* (pp. 25–36). Dordrecht: Springer.
- Dey, R., Pal, K., Bhatt, D., & Chauhan, S. (2004). Growth promotion and yield enhancement of peanut (*Arachis hypogaea* L.) by application of plant growthpromoting rhizobacteria. *Microbiological Research*, *159*, 371–394.
- Elias, F., Woyessa, D., & Muleta, D. (2016). Phosphate solubilization potential of rhizosphere fungi isolated from plants in Jimma Zone, Southwest Ethiopia. *International Journal of Microbiology*, *2016*, 1.
- Ellis, J. (2017). Can plant microbiome studies lead to effective biocontrol of plant diseases? *Molecular Plant–Microbe Interactions*. Available from <https://doi.org/10.1094/MPMI-12-16-0252-CR>.
- Enebe, M. C., & Babalola, O. O. (2020). Effects of inorganic and organic treatments on the microbial community of maize rhizosphere by a shotgun metagenomics approach. *Annals of Microbiology*, *70*, 49.
- Errakhi, R., Bouteau, F., Barakate, M., & Lebrihi, A. (2016). *Isolation and characterization of antibiotics produced by Streptomyces J-2 and their role in biocontrol of plant diseases, especially grey mould. Biocontrol of Major Grapevine Diseases* (pp. 76–83). Wallingford, UK: CAB International.
- Esitken, A., Pirlak, L., Turan, M., & Sahin, F. (2006). Effects of floral and foliar application of plant growth promoting rhizobacteria (PGPR) on yield, growth and nutrition of sweet cherry. *Scientia Horticulturae (Amsterdam)*, *110*, 324–327.
- Fakruddin, M., & Mannan, K. S. B. (2013). Methods for Analyzing Diversity of Microbial Communities in Natural Environments. *Ceylon Journal of Science (Biological Sciences)*, *42*(1), 19–33.
- Felczykowska, A., Bloch, S. K., Nejman-Falenczyk, B., & Baranska, S. (2012). Metagenomic approach in the investigation of new bioactive compounds in the marine environment. *Acta Biochimica Polonica*, *59*, 501–505.
- Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, *7*, 461–465.

- Gabor, E. M., Alkema, W. B., & Janssen, D. B. (2004). Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology*, 6, 879–886.
- Gamalero, E., Lingua, G., Tombolini, R., Avidano, L., Pivato, B., & Berta, G. (2005). Colonization of tomato root seedling by *Pseudomonas fluorescens* 92rkG5: spatio-temporal dynamics, localization, organization, viability, and culturability. *Microbial Ecology*, 50, 289–297.
- Guazzaroni, M., Alves, L., Lovate, G., Siqueira, G., Westmann, C., & Borelli, T. (2018). Metagenomic approaches for understanding new concepts in microbial science. *International Journal of Genomics*, Article ID 2312987.
- Gulati, A., Vyas, P., Rahi, P., & Kasana, R. C. (2009). Plant growth-promoting and rhizosphere-competent *Acinetobacter rhizosphaerae* strain BIHB 723 from the cold deserts of the Himalayas. *Current Microbiology*, 58, 371–377.
- Gutiérrez-Mañero, F. J., Ramos-Solano, B., Probanza, A. N., Mehouchi, J. R., Tadeo, F., & Talon, M. (2001). The plant-growth-promoting rhizobacteria *Bacillus pumilus* and *Bacillus licheniformis* produce high amounts of physiologically active gibberellins. *Physiologia Plantarum*, 111, 206–211.
- Haas, D., & Défago, G. (2005). Biological control of soil-borne pathogens by fluorescent pseudomonads. *Nature Reviews. Microbiology*, 3(4), 307–319.
- Halmann, J., Quadt-Hallmann, A., Mahaffee, W. F., & Kloepper, J. W. (1997). Bacterial endophytes in agricultural crops. *Canadian Journal of Microbiology*, 43, 895–914.
- Handelsman, et al. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, 5(10), R245–R249.
- Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, 68(4), 669–685.
- Harris, T., Buzby, P., Babcock, H., et al. (2008). Single molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, 320, 106–109.
- Hart, C., Lipson, D., Ozsolak, F., Raz, T., Steinmann, K., Thompson, J., et al. (2010). Single molecule sequencing: Sequence method to enable accurate quantitation. *Methods in Enzymology*, 472, 407–430.
- Hashem, A., Abd_Allah, E. F., Alqarawi, A. A., Radhakrishnan, R., & Kumar, A. (2017). Plant defense approach of *Bacillus subtilis* (Bera 71) against *Macrophomina phaseolina* (tassi) Goid in mung bean. *J Plant Interact*, 12, 390–401. Available from <https://doi.org/10.1080/17429145.2017.1373871>.
- Hayden, E. C. (2012). Nanopore genome sequencer makes its debut. *Nature*, 10051. Available from <https://doi.org/10.1038/nature.2012.10051>.
- Hiltner, L. (1904). Ueber neuere Erfahrungen und Probleme auf dem Gebiete der Bodenbakteriologie und unter besonderer Berücksichtigung der Grundung und Brache. *Arb. Deut. Landw. Gesell*, 98, 59–78.
- Hirsch, P. R., Miller, A. J., & Dennis, P. G. (2013). Do root exudates exert more influence on rhizosphere bacterial community structure than other rhizodeposits? In F. J. de Bruijn (Ed.), *Molecular Microbial Ecology of the Rhizosphere* (vol 1, pp. 229–242). Hoboken, New Jersey, USA: Wiley Blackwell.
- Iniguez, A. L., Dong, Y., & Triplett, E. W. (2004). Nitrogen fixation in wheat provided by *Klebsiella pneumoniae* 342. *Mol. Plant Microbe*, 17(10), 1078–1085.
- Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akesen, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, 12, 351–356.
- Kapri, A., & Tewari, L. (2010). Phosphate solubilization potential and phosphatase activity of rhizospheric *Trichoderma* spp. *Brazilian Journal of Microbiology*, 41, 787–795.
- Kaur, G., Sharma, P., Chhabra, D., Chand, K., & Mangat, G. S. (2017). Exploitation of endophytic *Pseudomonas* sp. for plant growth promotion and colonization in rice. *Journal of Applied and Natural Science*, 9(3), 1310–1316.
- Kour, D., Rana, K. L., Kaur, T., Sheikh, I., Yadav, A. N., Kumar, V., . . . Saxena, A. K. (2020). Microbe-mediated alleviation of drought stress and acquisition of phosphorus in great millet (*Sorghum bicolor* L.) by drought-adaptive and phosphorus-solubilizing microbes. *Biocatalysis and Agricultural Biotechnology*, 23, 101501.
- Kavino, M., Harish, S., Kumar, N., Saravanakumar, D., & Samiyappan, R. (2010). Effect of chitinolytic PGPR on growth, yield and physiological attributes of banana (*Musa* spp.) under field conditions. *Applied Soil Ecology*, 45, 71–77.
- Ko, K. C., Han, Y., Cheong, D. E., Choi, J. H., & Song, J. J. (2013). Strategy for screening metagenomic resources for exocellulase activity using a robotic, high-throughput screening system. *Journal of Microbiological Methods*, 94, 311–316.
- Kremer, R. J. (2019). Bioherbicides and nanotechnology: Current status and future trends. *Nano-Biopesticides Today and Future Perspectives*, 353–366.
- Kumar, S., Pandey, P., & Maheshwari, D. K. (2009). Reduction in dose of chemical fertilizers and growth enhancement of sesame (*Sesamum indicum* L.) with application of rhizospheric competent *Pseudomonas aeruginosa* LES4. *European Journal of Soil Biology*, 45(4), 334–340.
- Kumar, A., Gupta, A., Azooz, M. M., Sharma, S., Ahmad, P., & Dames, J. (2013). Genetic approaches to improve salinity tolerance in plants. In P. Ahmad, M. M. Azooz, & M. N. V. Prasad (Eds.), *Salt stress in plants*. New York: Springer.
- Kumar, A., Sharma, S., & Mishra, S. (2010). Influence of arbuscular mycorrhizal (AM) fungi and salinity on seedling growth, solute accumulation, and mycorrhizal dependency of *Jatropha curcas* L. *Journal of Plant Growth Regulation*, 29, 297–306. Available from <https://doi.org/10.1007/s00344-009-9136-1>.
- Kumari, S., Jha, A. K., Kumar, R., & Rajanikant. (2018). Role of Microbial Biotechnology in Sustainable Agriculture and Environment. *Chemical Science Review and Letters*, 7(25), 184–189.
- Leong, J. (1986). Siderophores: their biochemistry and possible role in the biocontrol of plant pathogens. *Annual Review of Phytopathology*, 24(1), 187–209.

- Li, X., Rui, J., Kiong, J., Li, J., He, Z., Anthony, C., & Yanncoull, M. R. (2014). Functional potential of soil microbial communities in maize rhizosphere. *PLoS One*, *9*, e112609.
- Lin, L., & Xu, X. (2013). Indole-3-acetic acid production by endophytic *Streptomyces* sp. En-1 isolated from medicinal plants. *Current Microbiology*, *67*(2), 209–217.
- Liu, L., Li, Y.-H., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, 251364. Available from <https://doi.org/10.1155/2012/251364>.
- Lorenz, P., & Eck, J. (2005). Metagenomics and industrial applications. *Nature Reviews. Microbiology*, *3*, 510–516.
- Majeed, A., Abbasi, M. K., Hameed, S., Imran, A., & Rahim, N. (2015). Isolation and characterization of plant growth-promoting rhizobacteria from wheat rhizosphere and their effect on plant growth promotion. *Frontiers in Microbiology*, *6*, 198.
- Mendes, R., Garbeva, P., & Raaijmakers, J. M. (2013). The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiology Reviews*, *37*, 634–663.
- Mendes, R., Krujic, M., de Bruijn, E., Dekkers, E., Van, D. E. R., voort, M., & Schneider, J. H. M. (2011). Deciphering the rhizosphere microbiomes for disease-suppressing bacteria. *Science (New York, N.Y.)*, *332*, 1097–1100. Available from <https://doi.org/10.1126/science.1203980>.
- Mittal, V., Singh, O., Nayyar, H., Kaur, J., & Tewari, R. (2008). Stimulatory effect of phosphate-solubilizing fungal strains (*Aspergillus awamori* and *Penicillium citrinum*) on the yield of chickpea (*Cicer arietinum* L. cv. GPF2). *Soil Biology & Biochemistry*, *40*, 718–727.
- Nair, H. P., Vincent, H., & Bhat, S. G. (2014). Evaluation of five in situ lysis protocols for PCR amenable metagenomic DNA from mangrove soils. *Biotechnology Reports*, *4*, 134–138.
- Niu, B., Paulson, J. N., Zheng, X., & Kolter, R. (2017). Simplified and representative bacterial community of maize roots. *Proceedings of the National Academy of Sciences*, *114*(12), E2450–E2459.
- Niu, S. Y., Yang, J., McDermaid, A., Zhao, J., Kang, Y., & Ma, Q. (2018). Bioinformatics tools for quantitative and functional metagenome and meta-transcriptome data analysis in microbes. *Briefings in Bioinformatics*, *19*(6), 1415–1429.
- Pankievicz, V., Amaral, F. P., Santos, K. F., Agtuca, B., Xu, Y., Schueller, M. J., . . . Pedrosa, F. O. (2015). Robust biological nitrogen fixation in a model grass–bacterial association. *The Plant Journal: for Cell and Molecular Biology*, *81*(6), 907–919.
- Paula, F., Rodrigues, J. L. M., Zhou, J. Z., Wu, L. Y., Mueller, R. C., Mirza, B. S., et al. (2014). Land use change alters functional gene diversity, composition and abundance in Amazon forest soil microbial communities. *Molecular Ecology*, *23*, 2988–2999.
- Prabha, R., Singh, DP, Gupta, S, Gupta, VK, El-Enshasy, HA, & Verma, MK (2019). *Rhizosphere metagenomics of Paspalum scrobiculatum l.(kodo millet) reveals rhizobiome multifunctionalities. Microorganisms*, *7*(12), 608.
- Quadt-Hallmann, A., Klopper, J., & Benhamou, N. (1997). Bacterial endophytes in cotton: mechanisms of entering the plant. *Canadian Journal of Microbiology*, *43*(6), 577–582.
- Quecine, M. C., Araujo, W. L., Tsui, S., Parra, J. R. P., Azevedo, J. L., & Pizzirani-Kleiner, A. A. (2014). Control of *Diatraea saccharalis* by the endophytic *Pantoea agglomerans* 33.1 expressing cry1Ac7. *Archives of Microbiology*, *196*, 227–234.
- Raaijmakers, J. M., Paulitz, T. C., Steinberg, C., et al. (2009). The rhizosphere: a playground and battlefield for soilborne pathogens and beneficial microorganisms. *Plant and Soil*, *321*, 341–361.
- Raaijmakers, J. M., Vlami, M., & De Souza, J. T. (2002). Antibiotic production by bacterial biocontrol agents. *Antonie Van Leeuwenhoek*, *81*(1), 537–547.
- Rajilić-Stojanović, M., Smidt, H., & De Vos, W. M. (2007). Diversity of the human gastrointestinal tract microbiota revisited. *Environmental Microbiology*, *9*, 2125–2136.
- Ramesh, R., Joshi, A., & Ghanekar, M. P. (2008). *Pseudomonads*: Major antagonistic endophytic bacteria to suppress bacterial wilt pathogen, *Ralstonia solanacearum* in the eggplant (*Solanum melongena* L.). *World Journal of Microbiology & Biotechnology*, *25*, 47–55.
- Rameshkumar, N., Krishnan, M., Kandeepan, C., & Kayalvizhi, N. (2014). Molecular and functional diversity of PGPR fluorescent *Pseudomonas* isolated from rhizosphere of rice (*Oryza sativa* L.). *International Journal of Advanced Biotechnology Research*, *5*, 490–505.
- Rashid, S., Charles, T. C., & Glick, B. R. (2012). Isolation and characterization of new plant growth-promoting bacterial endophytes. *Applied Soil Ecology*, *61*, 217–224.
- Ryu, C.-M., Shin, J. N., Qi, W., Ruhong, M., Kim, E. J., & Pan, J. G. (2011). Potential for augmentation of fruit quality by foliar application of bacilli spores on apple tree. *Plant Pathology Journal*, *27*, 164–169.
- Sabale, S., Suryawanshi, P., & Krishnaraj. (2019). Soil metagenomics: Concepts and applications. *Intechopen*. Available from <https://doi.org/10.5772/intechopen.88958>.
- Sandhya, V., Ali, S. Z., Grover, M., Reddy, G., & Venkateswarlu, B. (2010). Effect of plant growth promoting *Pseudomonas* spp. on compatible solutes, antioxidant status and plant growth of maize under drought stress. *Plant Growth Regulation*, *62*, 21–30.
- Sarkar, A., Ghosh, P. K., Pramanik, K., Mitra, S., Soren, T., Pandey, S., et al. (2018). A halotolerant Enterobacter sp. displaying ACC deaminase activity promotes rice seedling growth under salt stress. *Research in Microbiology*, *169*, 20–32.
- Schimel, J. P., & Schaeffer, S. M. (2012). Microbial control over carbon cycling in soil. *Frontiers in Microbiology*. Available from <https://doi.org/10.3389/fmicb.2012.00348>.
- Schoenfeld, T., Liles, M., Wommack, K. E., Polson, S. W., Godiska, R., & Mead, D. (2010). Functional viral metagenomics and the next generation of molecular tools. *Trends in Microbiology*, *18*, 20–29.
- Schreiter, S., Sandmann, M., Smalla, K., & Grosch, R. (2014). Soil type dependent rhizosphere competence and biocontrol of two bacterial inoculant strains and their effects on the rhizosphere microbial community of field-grown lettuce. *PLoS One*, *9*, e103726.
- Selvakumar, G., Kundu, S., Gupta, A. D., Shouche, Y. S., & Gupta, H. S. (2008). Isolation and characterization of nonrhizobial plant growth promoting bacteria from nodules of Kudzu (*Pueraria thunbergiana*) and their effect on wheat seedling growth. *Current Microbiology*, *56*, 134–139.

- Simonet, P., Nesme, J., Achouk, W., Agathos, S., et al. (2016). Back to the Future of Soil Metagenomics. *Frontiers in Microbiology*, 7, 73.
- Singh, R., Devi, T., Verma, V., & Rasool, S. (2014). Comparative studies on the extraction of metagenomic DNA from various soil and sediment samples of Jammu and Kashmir region in prospect for novel biocatalysts. *IOSR Journal of Environmental Science, Toxicology and Food Technology*, 8, 46–56.
- Singh, R. P., Johri, A. K., & Dua, M. (2020). Metagenomic analysis of microbial diversity in cotton rhizosphere soil in Alwar, India. *Microbiology Resource Announcements*, 9, e00987.
- Soni, R., Shaluja, B., & Goel, R. (2010). Bacterial community analysis using temporal gradient gel electrophoresis of 16 S rDNA PCR products of soil metagenomes. *Ekologija*, 56(3&4), 94–98.
- Steele, H. L., Jaeger, K. E., Daniel, R., & Streit, W. R. (2009). Advances in recovery of novel biocatalysts from metagenomes. *Journal of Molecular Microbiology and Biotechnology*, 16, 25–37.
- Suenaga, H., Koyama, Y., Miyakoshi, M., Miyazaki, R., Yano, H., Sota, M., et al. (2009a). Novel organization of aromatic degradation pathway genes in a microbial community as revealed by metagenomic analysis. *The ISME Journal*, 3, 1335–1348.
- Suman, A., Shasany, A. K., Singh, M., Shahi, H. N., Gaur, A., & Khanuja, S. P. S. (2001). Molecular assessment of diversity among endophytic diazotrophs isolated from subtropical Indian sugarcane. *World Journal of Microbiology and Biotechnology*, 17(1), 39–45. Available from <https://doi.org/10.1023/A:1016624701517>.
- Tanveer, A., Yadav, S., & Yadav, D. (2016). Comparative assessment of methods for metagenomic DNA isolation from soils of different crop growing fields. *3 Biotech*, 6, 220.
- Taulé, C., Mareque, C., Barlocco, C., Hackembruch, F., Reis, V. M., Sicardi, M., & Battistoni, F. (2012). The contribution of nitrogen fixation to sugarcane (*Saccharum officinarum* L.), and the identification and characterization of part of the associated diazotrophic bacterial community. *Plant and Soil*, 356, 35–49.
- Tottey, W., Denonfoux, J., Jaziri, F., Parisot, N., Missaoui, M., Hill, D., et al. (2013). The Human Gut Chip “HuGChip,” an Explorative Phylogenetic Microarray for Determining Gut Microbiome Diversity at Family Level. *PLoS One*, 8(5), e62544. Available from <https://doi.org/10.1371/journal.pone.0062544>.
- Upadhyay, S. K., Singh, D. P., & Saikia, R. (2009). Genetic diversity of plant growth promoting rhizobacteria isolated from rhizospheric soil of wheat under saline condition. *Current Microbiology*, 59, 489–496.
- Van Loon, L. C., Bakker, P. A., & Pieterse, C. M. (1998). Systemic resistance induced by rhizosphere bacteria. *Annual review of phytopathology*, 36(1), 453–483.
- Van Nostrand, J. D., He, Z., & Zhou, J. (2012). Use of functional gene arrays for elucidating in situ biodegradation. *Frontiers in Microbiology*, 3, 339.
- van Elsas, J. D., Chiurazzi, M., Mallon, C. A., Elhottová, D., Křišťáček, V., & Salles, J. F. (2012). Microbial diversity determines the invasion of soil by a bacterial pathogen. *Proceedings of the National Academy of Sciences*, 109(4), 1159–1164.
- Vardharajula, S., Zulfikar Ali, S., Grover, M., Reddy, G., & Bandi, V. (2011). Drought-tolerant plant growth promoting *Bacillus* spp.: effect on growth, osmolytes, and antioxidant status of maize under drought stress. *Journal of Plant Interaction*, 6, 1–14.
- Verma, P., Yadav, A. N., Khannam, K. S., Mishra, S., et al. (2019). Appraisal of diversity and functional attributes of thermotolerant wheat associated bacteria from the peninsular zone of India. *Saudi Journal of Biological Sciences*, 26(7), 1882–1895.
- Verma, P., Yadav, A. N., Khannam, K. S., Panjiar, N., Kumar, S., Saxena, A. K., et al. (2015). Assessment of genetic diversity and plant growth promoting attributes of psychrotolerant bacteria allied with wheat (*Triticum aestivum*) from the northern hills zone of India. *Annals of Microbiology*, 65, 1885–1899.
- Vessey, J. K. (2003). Plant growth promoting rhizobacteria as biofertilizers. *Plant and Soil*, 255, 571–586.
- Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., Van Elsas, J. D., et al. (2009). Terra genome: a consortium for the sequencing of a soil metagenome. *Nature Reviews. Microbiology*, 7, 252.
- Volossiouk, T., Robb, E. J., & Nazar, R. N. (1995). Direct DNA extraction for PCR-mediated assays of soil organisms. *Applied Environmental Microbiology*, 61, 3972–3976.
- Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D., & DeRisi, J. L. (2002). Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24), 15687–15692. Available from <https://doi.org/10.1073/pnas.242579699>, Nov 26 Epub 2002 Nov 12. PMID. Available from 12429852, PMCID: PMC137777.
- Wang, Z., Zhang, X.-X., Lu, X., Liu, B., Li, Y., Long, C., et al. (2014). Abundance and Diversity of Bacterial Nitrifiers and Denitrifiers and Their Functional Genes in Tannery Wastewater Treatment Plants Revealed by High-Throughput Sequencing. *PLoS One*, 9(11), e113603. Available from <https://doi.org/10.1371/journal.pone.0113603>.
- Wang, X., Wang, C., Sui, J., Liu, Z., Li, Q., Ji, C., et al. (2018). Isolation and characterization of phosphofungi, and screening of their plant growth-promoting activities. *AMB Express*, 8.
- Weller, D. M., Raaijmakers, J. M., Gardener, B. B., & Thomashow, L. S. (2002). Microbial populations responsible for specific soil suppressiveness to plant pathogens. *Annual Review of Phytopathology*, 40, 309–348.
- Wu, L., Thompson, D. K., Li, G., Hurt, R. A., Tiedje, J. M., & Zhou, J. (2001). Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Applied and Environmental Microbiology*, 67, 5780–5790.
- Yadav, A., Kumar, V., Dhaliwal, H., Prasad, R., & Saxena, A. (2018). *Microbiome in Crops: Diversity, Distribution, and Potential Role in Crop Improvement. Crop Improvement through Microbial Biotechnology* (pp. 305–322). Elsevier, chapter 15.
- Yadav, N., Ghimire, S., Shrestha, S., Sah, B., Sarker, A., & Sah, S. (2017). Source of resistant against *Fusarium* wilt and *Stemphylium* blight in lentil (*Lens culinaris* Medikus). *International Journal of Applied Sciences and Biotechnology*, 5(1), 102–107. Available from <https://doi.org/10.3126/ijasbt.v5i1.17027>.

- Yang, J., Kloepper, J. W., & Ryu, C. M. (2009). Rhizosphere bacteria help plants tolerate abiotic stress. *Trends in Plant Science*, 14(1), 1–4.
- Yadav, A. N. (2020). *Plant microbiomes for sustainable agriculture: current research and future challenges. Plant microbiomes for sustainable agriculture* (pp. 475–482). Springer.
- Yildirim, E., Karlidag, H., Turan, M., Dursun, A., & Goktepe, F. (2011). Growth, nutrient uptake, and yield promotion of broccoli by plant growth promoting rhizobacteria with manure. *Horticultural Science*, 46, 932–936.
- Yuan, Z., Li, R., Pang, Z., Zhou, Y., Fallah, N., SSHu, C., & Lin, W. (2020). Metagenomic analysis exploring taxonomic and functional diversity of soil microbial communities in sugarcane fields applied with organic fertilizer. *Hindwani BioMed Research International*, Article ID 9381506.
- Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38, 95–109.
- Zhang, Y., Chen, F.-S., Wu, X.-Q., Luan, F.-G., Zhang, L.-P., Fang, X.-M., et al. (2018). Isolation and characterization of two phosphate-solubilizing fungi from rhizosphere soil of moso bamboo and their functional capacities when exposed to different phosphorus sources and pH environments. *PLoS One*, 13.
- Zhao, J, Sun, Q, Quentin, M, Ling, J, Abad, P, Zhang, X, ... Xie, B (2021). A Meloidogyne incognita C-type lectin effector targets plant catalases to promote parasitism. *New Phytologist*, 232(5), 2124–2137.
- Zhou, J., Bruns, M. A., & Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology*, 62, 316–322.
- Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology: A technology review and future perspective. *Sciences China Life Sciences*, 53, 44–57.

Concepts and applications of bioinformatics for sustainable agriculture

Ezgi Çabuk Şahin¹, Yıldız Aydın¹, Tijis Gilles², Ahu Altınkut Uncuoğlu³ and Stuart J. Lucas²

¹Department of Biology, Faculty of Science & Arts, Marmara University, Istanbul, Turkey, ²Sabancı University Nanotechnology Research and Application Center, Istanbul, Turkey, ³Department of Bioengineering, Faculty of Engineering, Marmara University, Istanbul, Turkey

28.1 Introduction—a conceptual framework for sustainable agriculture

Over the last few years, “sustainable agriculture” has become one of the key concepts underpinning global policy for agricultural development and food security. Defined by the Food and Agriculture Organization of the United Nations as “agriculture that meets the needs of present and future generations, while ensuring profitability, environmental health and social and economic equity” (FAO, 2022), this concept recognizes that current agricultural practices rely on the consumption of finite resources and production of damaging waste more rapidly than the global ecosystem can regenerate them. Accordingly, on the formal adoption by the world leaders of the UN Sustainable Development Goals in September 2015, promoting sustainable agriculture was included as a vital pillar of SDG 2 (Zero Hunger).

Measuring and achieving sustainability in agriculture is far from simple. Positive actions to protect genetic diversity, reduced and more efficient use of natural resources and agrochemicals, minimization and valorization of waste products, and the restructuring of trade and distribution networks for agricultural products are all essential. One major paradigm shift toward sustainability is a move away from elite cultivars of crops, which provide maximized yields at the expense of high agricultural inputs and limited genetic diversity, toward diverse locally adapted crops that are able to flourish and provide reliable yields with fewer inputs and suboptimal environmental conditions (Shelef, Weisberg, & Provenza, 2017). However, the performance of such varieties depends on a complex interplay of genotype, phenotype, and environmental interactions (Fig. 28.1).

It is in this complexity that bioinformatics is increasingly important for agricultural science. In the genotype sphere, rapid technological developments in high-throughput DNA sequencing and genotyping methods mean that the amount of genetic data available for crop species is growing exponentially (Scheben, Batley, & Edwards, 2018). Phenotyping for many traits remains time-consuming and labor-intensive, but techniques utilizing autonomous vehicles or scanning platforms, spectroscopy at multiple bandwidths, and high-resolution imaging are also starting to be deployed in the field (Shakoor, Lee, & Mockler, 2017). Finally, in the environment sphere, weather stations collect continuous climatic data and remote sensors operating from aerial vehicles and satellites can assess soil and field conditions down to square meter resolution (Shafi et al., 2019). All of these emerging technologies generate enormous datasets; mining these data for functionally significant features and elucidating the relationships between them are the major challenges facing agricultural bioinformaticians.

In this chapter, we focus primarily on the bioinformatic tools and resources currently available for exploiting genetic data from crop and livestock species, and the gaps that must be filled in order to convert these resources into improvements in agricultural sustainability.

28.2 Database resources for agricultural bioinformatics

The proliferation of nucleotide sequences and related biological data produced by high-throughput sequencing and other “omics” platforms in recent years has also led to a proliferation of biological databases. These databases collect data from diverse research studies in a consistent format and typically include integrated software tools designed for the

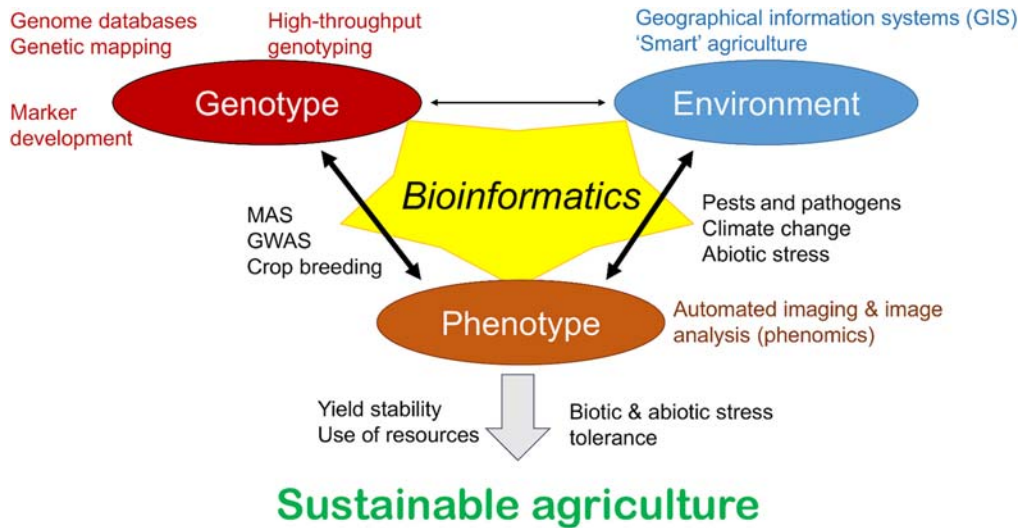


FIGURE 28.1 Genotype, environment and phenotype are all inter-related in developing sustainable agricultural practices; modern technologies generate huge datasets analyzing all three areas, which need to be interpreted by bioinformatics.

storage of biological information and to update the records stored in the system. However, no single database can currently integrate all of the different kinds of “big data” that are relevant to agriculture. There is also a fairly rapid turnover of more specialized databases that are funded as part of a specific research program but may not be maintained beyond the end of the initial grant. In these circumstances, finding the most recent data and software tools to address a particular scientific question can be a bewildering challenge.

In [Table 28.1](#), we present a nonexhaustive list of current databases that are actively updated, cover multiple crop species, and provide access to data that is useful to bioinformaticians. For example, the first port of call for nearly any scientist exploring genetic data are the members of the International Nucleotide Sequence Database Collaboration, or INSDC: NCBI-GenBank, EMBL-EBI, and DDBJ. The first two of these databases were initiated in 1982, and since 1987 all three databases have been committed to ensuring that publically available nucleotide sequence data is preserved and made accessible to users across the world ([Arita, Karsch-mizrachi, & Cochrane, 2020](#)). In particular, open scientific data policies most recently codified in the FAIR ([Wilkinson et al., 2016](#)) (Findable, Accessible, Interoperable and Renewable) principles require that as a minimum, raw sequence data from newly published studies are available in the INSDC sequence archives, and in many cases assembled and annotated data are also deposited. Each of these databases mirrors the same data collection, but each also include supercomputer infrastructure and unique tools allowing users to interact with the data in different ways.

One of the main goals of bioinformatics is to understand the relationship between the sequences of nucleotides or amino acids, the three-dimensional structures they form, and the molecular functions of these structures. Unfortunately, the relationship between sequence and structure is not trivial, so most known protein structures have been resolved empirically or by modeling based on a known structure with a related sequence. The single comprehensive repository of protein, nucleic acid, and complex 3D structures is the worldwide Protein Data Bank (wwPDB), which is celebrating its 50th anniversary in 2021 and now contains >170,000 entries ([Berman, Henrick, & Nakamura, 2003](#)).

These global databases are an essential resource for any study aiming to mine and reevaluate existing datasets. However, the resources required to store such a huge amount of data (Genbank exceeded 9 Petabytes = 9 million GB in 2020) and their breadth of scope (covering sequences from all forms of life) means that these databases do not incorporate some types of analysis that are particularly valuable for crop genomics. For example, homology and colinearity of genes between related species is common throughout the plant kingdom and an extremely powerful tool for inferring the function and evolutionary history of genes in nonmodel crops, for which experimental data may be limited ([Tello-Ruiz et al., 2018](#)). Several *comparative genomics databases* (Gramene, Ensembl-Plants, Phytozome, and PLAZA) exist to address this need, with slightly different datasets and emphases. Each of these databases contain fully annotated and curated genomes representing diverse plant species, which can be of considerable value in identifying orthologous genes and identifying species-specific mutations.

In many cases, the newest genome assemblies and annotations for a given crop are not found in the completely publically accessible databases mentioned earlier, because they are covered by a limited early release agreement (which

TABLE 28.1 Summary of current online database resources for crop-related functional genetics.

DB name	Current version/date	Contents	URL and references
Global genetics databases			
INSDC databases: DDBJEMBL-EBINNCBI	SRA/ENA:updated dailyGenBank: release 248, February 2022 (update each 2 months)	The three partner databases in the International Nucleotide Sequence Database Consortium are the global repository for nucleotide sequencing data from all species (Sequence Read Archive/European Nucleotide Archive). Project metadata (BioProject and BioSample) and raw sequence data are mirrored between all three databases. These databases also give access to the largest worldwide repository of annotated DNA sequences (GenBank).	https://www.ddbj.nig.ac.jp/index-e.html https://www.ebi.ac.uk/ena/browser/ https://www.ncbi.nlm.nih.gov/genbank/ Arita et al. (2020)
wwPDB	Updated weekly	Worldwide Protein Data Bank. Core archive of 3D structures from proteins, nucleotides, and complexes relevant to all aspects of biomedicine and agriculture.	http://www.wwpdb.org Berman et al. (2003)
Comparative genomics databases			
Gramene	Release 64, October 2021	Curated open-source integrated database of fully annotated genes and genomes from 93 plant species. Incorporates the Plant Reactome pathway database.	http://www.gramene.org/ Tello-Ruiz et al. (2018)
Ensembl-Plants	Release 49, December 2020	Provides access, search, and comparison tools for a range of data types anchored to annotated reference genomes from 90 plant species.	https://plants.ensembl.org/index.html Howe et al. (2020)
Phytozome	v13	Provides data access and search tools for mining 93 annotated genomes representing 82 plant species. Through sequence comparison families of related genes representing the modern descendants of ancestral genes are constructed at key phylogenetic nodes.	https://phytozome.jgi.doe.gov/pz/portal.html Goodstein et al. (2012)
PLAZA	v5.0, 2021	Comparative genomic database integrating and annotating sequence data from 39 different plant genomes, and evolutionary relationships between them.	https://bioinformatics.psb.ugent.be/plaza/ Van Bel et al. (2018)
Taxon-specific databases			
CGD	2019	Citrus Genome Database. Access to 11 citrus genomes, along with details of genes, markers, quantitative trait loci (QTLs), and genetic maps from 67 species.	https://www.citrusgenomedb.org/
CuGenDB	December 2019	Cucurbit Genomics Database. Provides tools for browsing 16 genomes from 10 species, and comparison between genomes and with RNA-seq datasets.	http://cucurbitgenomics.org/ Zheng et al. (2019)
GDR	2019	Genome Database for Rosaceae. Includes genome, genotype, phenotype, and molecular marker data for 10 Rosaceae crop species and wild relatives.	https://www.rosaceae.org/ Jung et al. (2019)
GrainGenes	2020	Gives access to genome data, expressed sequence tag, germplasm, marker, and QTL information for wheat, barley, rye, and oat.	https://wheat.pw.usda.gov/GG3/ Blake et al. (2019)

(Continued)

TABLE 28.1 (Continued)

DB name	Current version/date	Contents	URL and references
HWG	2020	Hardwood Genomics Project—Open-source database for genome, transcriptome, and molecular marker data for forest trees, including several crop species.	https://www.hardwoodgenomics.org/ Kremer et al. (2012)
LegumeIP	V3, 2020	Genome data from 22 species, RNA-seq data, and tools for comparative and translational genomics.	http://plantgrn.noble.org/LegumeIP/gdp/ Li, Dai, Zhuang, and Zhao (2016)
Biological function-focused databases			
plaBiPD (successor of GabiPD)	July 2019	Provides searchable and browseable access to seven crop genomes as well as <i>Arabidopsis thaliana</i> . Also access to the MapMan/Mercator plant functional annotation tools, and a list of all published plant genome sequences.	https://plabipd.de/index.ep Usadel, Schwacke, Nagel, and Kersten (2012)
PRGdb	3.0, 2017	Open resource for confirmed and predicted Pathogen Receptor Genes (PRG), currently with representatives from 268 plant species.	http://prgdb.org/prgdb/ Osuna-Cruz et al. (2018)
PlantsDB (PGSB-PlantsDB)		Database of 12 plant genomes and browsing tools, along with specific databases of drought stress genes (DroughtDB) and plant repetitive elements (PGSB-REdat).	http://pgsb.helmholtz-muenchen.de/plant/plantsdb.jsp Spannagl et al. (2016)
PceRBase	2018	Contains ncRNA sequence data from 28 plant species and predicts competing endogenous RNA (ceRNAs) that act as decoys preventing sRNA-mediated gene regulation.	http://bis.zju.edu.cn/pcernadb/index.jsp Yuan et al. (2017)
PLaMoM		Curated, searchable database of Plant Mobile Macromolecules, including mRNAs, small RNAs, and proteins, known to be transported between plant cells.	http://www.byanbioinfo.org/plamom/ Guan et al. (2017)
Plant DNA C-values database	Release 7.1, April 2019	Manually curated database of nuclear DNA content, currently covering 12,273 plant species, ranging from algae to higher plants.	https://cvalues.science.kew.org/ Pellicer and Leitch (2020)
PlantPAN	V 3.0, 2019	Plant Promoter Analysis Navigator contains transcription factor-binding site information for seven model and crop plant genomes, and tools for scanning user-defined promoters.	http://plantpan.itps.ncku.edu.tw/index.html Chow et al. (2019)
Plant rDNA database	Release 3.0, March 2017	Compiles data concerning the number, position, and structure of ribosomal DNA genes from 2148 plant species, and also telomere sequences where available.	https://www.plantrdnadatabase.com/

typically allows users to carry out studies based on individual genes but not genome-wide analyses) and/or their annotation is still being finalized. In addition, many other types of nongenomic data are extremely important in crop research, including genetic maps, quantitative trait loci (QTLs), molecular marker data, and population genetic studies. These types of data are most likely to be found in *taxon-specific databases* that have been established to serve the specific needs of researchers working on a particular class of plants, such as cereals, legumes, or forest trees (Table 28.1). Similar resources have also been developed focusing on specific research initiatives rather than taxons, such as the URGI platform that supports the diverse projects of INRAE in France (<https://urgi.versailles.inra.fr/>). Studies targeting crop genetic improvement should not overlook these databases, to ensure that they take into account the most current genome annotations and functional data.

A final category of databases is those focused on a specific *biological function* across plant species. Many of these address traits that are of particular interest for sustainable agriculture, such as the Pathogen Receptor Genes detailed in PRGdb (Osuna-Cruz et al., 2018), and the drought-stress-responsive genes reported in PlantsDB (Spannagl et al., 2016). Others provide insight into cutting-edge areas of plant biological research for which the practical applications are not yet known, but that may provide crucial future improvements in sustainable crop production, such as the intracellular movement of macromolecules (Guan et al., 2017) and the activity of competing endogenous RNAs (Yuan et al., 2017).

In summary, an enormous amount and variety of genomic data is already available for crop bioinformatics, and much of it has not yet been analyzed from the perspective of improving sustainability. Appropriate utilization of these resources should be a priority for future studies. However, in order to employ these data effectively, we also need to consider the ways in which loci from individual crop genome sequences are related to field crop populations, by means of molecular markers and genetic maps.

28.3 Genome mapping

Genetic maps comprised a linear ordering of molecular or morphological markers along “linkage groups.” The alleles found within a linkage group change depending on frequencies of crossover or recombination during meiosis (Meksem & Kahl, 2005). As this frequency varies along chromosomes, the genetic map is not proportional to the physical location of genes or their separation in base pairs. Instead, genetic mapping is a statistical method that uses Mendelian principles of segregation and recombination to correlate genes’ localizations and functions. The “map unit” is the centiMorgan (cM); 1 cM corresponds to a 1% probability of recombination occurring between two genetic loci within a linkage group (Schneider, 2005). An ideal genetic map includes as many linkage groups as the haploid chromosome number of a species. In addition, it is desirable that there is no gap larger than 20 cM between pairs of markers (de Vienne, 2002). “High-resolution” genetic maps may be generated by screening a large number of genetic markers, to further reduce the gap sizes.

The first genetic mapping was carried out by Morgan and Sturtevant in 1911 on the gender character of a F_2 fruit fly population based on fragmentation and recombination (Semagn, Bjørnstad, & Ndjiondjop, 2006). There are three basic principles underlying genetic mapping: (1) recombination is a random process, so the further apart two loci are, the more often it occurs between them; (2) as a result, neighboring genes on the chromosome are almost always inherited together; (3) therefore if two phenotypic or genetic markers are always coinherited in new generations, their genes are located close to each other. As genetic distance is a statistical value, a population of related individuals is required to produce the genetic map. Creating a genetic map requires selection of the most suitable mapping population, DNA isolation, an appropriate method of screening and scoring a specific set of markers in each isolated DNA sample, and statistical analysis of the results (Boopathi, 2013).

28.3.1 Molecular marker systems and populations used for genetic mapping

The degree of distinction between the parents of a population to be studied is a primary consideration in genetic mapping. To be “informative” in a given population, a genetic marker must be “polymorphic,” revealing different types (alleles) in their respective parents. Genetic markers are based on differences in DNA, being the inherited molecule; the genetic code has been revealed and can be interrogated using a variety of molecular tools that act on specific DNA sequences (e.g., restriction enzymes and polymerases). Molecular markers are used widely in agricultural genetics and are foundational for modern plant and animal systematics, breeding, and evaluation of gene resources. A large number of DNA marker types have been developed as molecular markers for use in genetic mapping studies. The first DNA markers used in agriculture relied on digestion of DNA with restriction enzymes followed by hybridization of the fragments to a radioactive probe to detect “restriction fragment length polymorphisms (RFLP)” (Botstein, White, Skolnick, & Davis, 1980). Genetic maps using RFLP molecular markers were created by Helentjaris, Slocum, Wright, Schaefer, and Nienhuis (1986) initially in maize and tomato plants, followed by similar studies in many other crops.

Compared to hybridization, the speed and convenience of the polymerase chain reaction (PCR) technique developed in the late 1980s led to a proliferation of amplification-based marker types. Many of these use primer sequences that appear at many sites in a genome, giving multiple potential polymorphic fragments from a single reaction, such as AFLP (amplified fragment length polymorphism) (Vos et al., 1995), RAPD (random amplified polymorphic DNA) (Michelmore, Paran, & Kesseli, 1991), SSR (simple sequence repeat) (Akkaya, Bhagwat, & Cregan, 1992), ISSR (inter simple sequence repeat) (Zietkiewicz, Rafalski, & Labuda, 1994), and SRAP (sequence related amplified polymorphism) (Li & Quiros, 2001). Others use nucleotide sequencing data to amplify a specific genetic locus, such as STS

(sequence tagged site) (Olson, Hood, Cantor, & Botstein, 1989), EST (expressed sequence tag), and SCAR (sequence characterize amplified region) (Paran & Michelmore, 1993). Still others combine known sequences with one of the more promiscuous methods mentioned above to obtain the advantages of both, such as CAPS (cleaved amplified polymorphic sequence) (Konieczny & Ausubel, 1993), EST-SSR (Cordeiro, Casu, McIntyre, Manners, & Henry, 2001), and TRAP (target region amplified polymorphism) (Hu & Vick, 2003). Since the development of high-throughput sequencing platforms, there has been a shift toward single-nucleotide polymorphism (SNP) markers, as these are ubiquitous and frequent (at least 1 per kbp in most genomes). The most common of these marker types are described in more detail in Section 28.4.1.

Population selection and development is also crucial for successful genetic mapping. Most mapping populations start from a single cross between two parents; in order to be informative, these should differ from each other as much as possible in the trait or traits of interest. From this initial cross, many different kinds of progeny lines, such as F_1 , F_2 , RIL (recombinant inbred lines), BC (backcross), DH (double haploids), and more can be developed (Meksem & Kahl, 2005). The choice of population structure is generally a trade-off between increasing map resolution—as each generation of progeny allows more recombination events to take place—and the time and expense necessary to produce the lines.

F₁ population: The F_1 population is the immediate progeny of crossing two suitable parents. As a result, F_1 hybrids should be heterozygous at all loci where the parents differ, assuming that the parental lines were homozygous. For this reason, and the limited number of recombination events produced in a single generation, they are not used for mapping in most crops. However, in species for which homozygous parental lines cannot be generated from a species due to self-incompatibility, inbreeding depression, or prohibitively long generation time (such as some tree crops), heterozygous parental plants are used to derive F_1 mapping populations (Zhigunov et al., 2017).

F₂ population: F_2 plants are the simplest form of segregating population, produced by selfing or crossing the F_1 hybrids. Loci that were heterozygous (AB) in both F_1 parents segregate in their F_2 progeny as AA:AB:BB in the ratio 1:2:1. The increase in the number both of homozygous loci and recombination events make F_2 populations more powerful than F_1 for genetic mapping. However, F_2 lines cannot be maintained indefinitely; their heterozygous loci segregate further in their offspring, meaning that inbred F_3 progeny from the same F_2 line are not genetically identical to each other. F_2 populations are often the most practical choice for genetic mapping in livestock (Falker-Gieske et al., 2019).

RILs: These lines are formed by inbreeding of selected F_2 individuals through multiple generations (down to F_6 – F_9) using only one seed from each line to continue it in the next generation (single seed descent). Each successive inbred generation increases the proportion of homozygous loci, reaching >99% homozygosity by the F_9 generation. Therefore a population of RILs includes all the genetic variation present in the F_2 population, but as stable lines that will remain more or less genetically fixed as long as they are inbred. Another advantage of RILs is that additional recombination events at each generation increase their resolution for linkage analysis. Therefore RILs are the preferred genetic mapping population in most inbreeding crop species, with the disadvantage that it can take 3–4 years to produce the lines (Barh, Khan, & Davies, 2015).

BC: Where the aim is to eliminate many of the negative characteristics of one parent in favor of the other (e.g., introgressing a disease resistance gene from a nondomesticated relative into an elite crop variety), it is common to “backcross” F_1 hybrids with the preferred parental line, instead of self or intercrossing within the F_1 generation. The resulting BC_1F_1 population may either be selfed to generate RILs as above, or repeatedly backcrossed to the same parent to produce BC inbred lines (BILs). Alternatively, a RIL line carrying a desirable trait locus may be backcrossed in the same way, using genetic markers to select progeny that includes the trait locus. This produces near isogenic lines (NILs) that are useful for fine-mapping of a specific locus, with a largely constant genetic background (Kooke, Wijnker, & Keurentjes, 2012).

DH: Doubled haploids are generated by rescue of haploid gametes from seeds, followed by chemically induced chromosome doubling. As the two chromosomes are identical, this can produce stable, 100% homozygous lines from F_1 hybrids in a single generation, rapidly creating a permanent resource for mapping (Meksem & Kahl, 2005). However, not all crop species or even varieties of the same species are amenable to the tissue culture needed to produce DH populations.

Exotic populations: while biparental populations are the backbone of genetic mapping and breeding in crops, they only include the genetic diversity contained in the two parental lines. Therefore the resulting genetic maps give poor resolution in regions where there is limited recombination between the parents, and may not be relevant to varieties that are not closely related to either parent. Therefore in the last decade more complex populations have been developed for some crops using a larger number of parents, such as MAGIC (multiparent advanced intercross) and NAM (nested association mapping) (Scott et al., 2020). Developing these populations requires considerable investment, but they show promise for mapping complex multigene traits such as abiotic stress resistance.

The choice of mapping population also depends on the marker system used. “Dominant” markers (such as the individual bands produced by RAPD primers) only show whether an allele is present or absent, so they cannot distinguish between homozygous and heterozygous loci. “Codominant” markers are usually preferred, as they give a different signal for each parental allele, both of which are observed in heterozygotes. However, in highly homozygous lines such as RILs and DH, dominant markers are equally informative (Ferreira, da Silva, da Costa e Silva, & Cruz, 2006).

After determining marker polymorphisms in the chosen population, the final stage of genetic mapping is linkage analysis. This is a complex statistical process that requires knowledge of the population structure and understanding of the biology directing genetic recombination (Semagn et al., 2006), in order to calculate recombination rates and map distances. Therefore although software tools already exist for calculating genetic maps from marker data (Cheema & Dicks, 2009), there is still a need for bioinformaticians to develop improved models, especially for more complex population structures (Meksem & Kahl, 2005).

28.3.2 Genetic mapping, physical mapping, and genome sequencing

The arrangement of genes and DNA markers on a chromosome is demonstrated by both genetic and physical maps. While recombination frequencies determine the distances between locations on genetic map, physical maps depend on measurements of the amount of DNA between loci to determine their proximity. For example, methods such as radiation, enzymes, or shear forces are used to break DNA molecules randomly into large fragments, which can be preserved as bacterial or yeast artificial chromosomes (BACs = 100–350 kb, YACs = 100 kb–2 Mb). These are then screened to determine which DNA markers co-occur on the same chromosome fragments, while overlapping fragments are compared to determine the marker order. Cytogenetic mapping is another type of physical map that provides direct visualization of DNA landmarks on microscopic chromosomes, calculating the distance between them. Optical mapping increases the resolution of this approach, using microscopic observation of specific labeled sequence motifs on linearized DNA fragments (Levy-Sakin & Ebenstein, 2013). Physical mapping is less subject to variation between different genomic regions, populations, or environments than genetic mapping and typically provides a higher resolution (Paterson, 2009); however, the molecular techniques involved require specialist equipment and expertise.

The physical map at its highest resolution is a complete genome sequence. Therefore, with the reducing costs of high-throughput sequencing, in some cases genome sequencing can remove the need for a physical map. However, in practice both physical and genetic maps are still a valuable resource for understanding the organization of the genome, as they provide long-range information about genome organization and structure over greater distances than sequencing reads. Physical and genetic maps can unravel the complexity of highly repetitive or polyploid genomes that are intractable to sequence assembly, while high-resolution maps provide the scaffold on which contiguous sequences (contigs) are ordered to produce a whole chromosome sequence (O'Rourke, 2014). Maps act as a bridge between breeding and genome research for map-based cloning and marker-assisted selection (MAS). For example, genome sequences are an excellent resource for defining new candidate molecular markers to improve mapping resolution and identifying candidate genes within a trait locus. On the other hand, comparison of the location and order of markers between different maps can provide important insights into genome structural rearrangements and evolutionary relationships of even distant individuals. Furthermore, genetic mapping directly tests the transition of genes from parent to progeny, a core aspect of crop improvement (Yu & Main, 2015).

In summary, although genome sequencing is greatly increasing our knowledge of the genetics of a broad variety of species, the complementary information provided by genetic and physical maps remains invaluable for cloning candidate genes and developing molecular breeding systems (Varshney et al., 2014). Therefore effective bioinformatic tools are essential for integrating these different kinds of data.

28.3.3 Comparative mapping

One way in which computational methods have been applied to agriculture is in comparative mapping, using map details from one species (either genetic or physical, including DNA sequence) to deduce possible gene structure in another species. This method uses specific knowledge, such as the entirely sequenced gene of a model plant species, to deduce the possible gene structure in the genome of an “orphan crop” for which DNA-level knowledge is missing (Paterson, 2009).

The development of DNA markers not only allowed the rapid generation of comprehensive genetic maps of agricultural species but also made interspecies comparisons possible. When conserved molecular markers are mapped across related species, it is possible to align the chromosomes of those species to create comparative linkage maps. In this

way, genomic similarities between species are revealed so that genetic information about one species may be extended to others and evolutionary inferences can be drawn (Boopathi & Boopathi, 2013). Since a comparatively limited number of chromosomal breaks occurred during the radiation of mammals and of many crop families, gene order is typically maintained between similar species over large chromosomal segments. DNA sequence homology can detect orthologous genes, and sets of these genes that share a common linear order (synteny) in two or more organisms are used to classify preserved genome segments and ancient chromosomal breakpoints. As mapping and sequencing efforts advance, the detection of smaller homologous chromosome segments is becoming possible, and comprehensive comparative maps are being established between numerous species. There are now fairly complex gene-based comparative maps between the genomes of humans, mouse, and rats and also among many mammalian species of agricultural significance (White & Matise, 2001). Comparative genetic mapping experiments in plants have also been performed on less characterized members of the Solanaceae, Poaceae, and Brassicaceae families, among others (Schmidt, 2000).

28.3.4 Practical applications of genetic mapping

Knowledge of plant and animal genetic maps has contributed to the production of agricultural crops and animals that are more nutritious, more sustainable, and more tolerant to drought, pests, and diseases. In the field of plant and animal breeding, genetic mapping is still the most valuable approach to identifying the genetic factors that underlie particularly quantitatively inherited traits (Meksem & Kahl, 2005).

Genetic maps are created to reveal how genes are arranged in chromosomes and can provide information about the chromosome structure and gene evolution of species. However, their most important application is in selection and identification of genes conferring stress resistance or other important traits such as seed quality and productivity, which can then be used in breeding through MAS. Labeling genes whose phenotypic characteristics are strongly linked to a molecular marker allows the producer to make indirect selection among seedlings by DNA screening, to check whether the marker is present. Although the cost per plant of DNA marker screening is usually higher than making phenotypic observations, if the phenotype is only expressed in later stages of the plant life cycle or under specific stress conditions, marker screening may save a lot of time and expense compared to traditional methods, making locus discovery and marker genotyping economical (Schneider, 2005). Therefore MAS based on genetic linkage maps shortens the generation time for breeders, thereby providing the opportunity to accelerate genetic quality improvement for producers. MAS enables the collection of genes in a single individual that provide resistance to diverse diseases and pests, high yield, and quality and facilitates the transfer of these genes to other individuals based on their associated markers, thus improved plants with resistance to adverse conditions can be obtained (Hayward, Tollenaere, Dalton-Morgan, & Batley, 2015).

MAS is also indispensable for breeding sustainable crops in countries with limited facilities and resources, because this technique allows selection without the need for large-scale field or greenhouse tests (Fang, Zhu, Wang, & Shangguan, 2016). Thus in a central laboratory, breeders can carry out breeding studies using the same technology for a wide variety of plants and characters. Molecular marker-supported breeding techniques have been widely used in crop breeding and developing many new crop cultivars and lines. For example, Jena and Mackill (2008) and Shi et al. (2009) implemented marker-assisted breeding strategies to simultaneously select for excellent crop quality characteristics along with rice blast resistance in rice, and tolerance to mosaic virus in soybean.

The first goal of genetic mapping is to develop molecular markers tightly linked to the trait of interest, and, second, the long-term goal is to clone genes that control traits based on these maps (Roose, 2007). Map-based cloning is a technology that is based on DNA markers and genetic relationships. It uses the linkage between target genes and molecular markers by scanning a genomic library (e.g., a BAC library) for the markers, allowing the cloning of the large fragments containing the target gene. Map-based cloning methods have been successfully used in the isolation and cloning of excellent agricultural genes for growth, development, and resistance in many species. For instance, salt stress and insect resistance genes in rice were successfully cloned by Tamura et al. (2014), while the main enzymes in the terpenoid metabolic pathway in maize were cloned by Lu et al. (2012). For crops where a high-quality genome sequence is available, it is possible to accelerate map-based cloning, by identifying candidate genes located close to molecular markers *in silico* prior to functional validation *in planta*.

28.4 DNA marker development and application to genotyping

The production and application of informative genetic maps described in the previous section relies on the ability to develop polymorphic DNA markers as efficiently as possible. The genome resources mentioned in Section 28.2 provide an opportunity for researchers to rapidly identify sequence polymorphisms within a crop population using

bioinformatics, provided that they have access to both a reference genome and sequencing data from individuals belonging to the population of interest. For less studied crops, markers that do not require prior knowledge of the genome sequence may be preferable. In this section, we consider the relative advantages and disadvantages of each marker type from this perspective, followed by the genotyping technologies that are used to screen these markers. Finally, we give a case study of marker development to map a multigene trait in a complex genome.

28.4.1 DNA marker types, their advantages and disadvantages

28.4.1.1 Restriction fragment length polymorphism

RFLP genotyping technology was initially applied in human genetic research to create a human marker map in the 1970s (Botstein et al., 1980) and later adopted in plant breeding programs (Beckmann & Soller, 1986). It is based on the genetic variability between germplasm lines at specific restriction endonuclease sites, which results in different length fragments when genomic DNA is cleaved with these enzymes. Length variations were visualized by separation of the DNA fragments by gel electrophoresis, and binding a labeled probe to the separated DNA. In plant species a lot of variation was found in the DNA annealing sites of restriction endonucleases. RFLP markers made it possible for the first time to identify markers closely linked to single-gene or more complex genetic traits, to transfer these traits into elite breeding germplasm, and accelerate breeding by identifying plants homozygous or heterozygous for the preferred genetic allele (Tanksley, Young, Paterson, & Bonierbale, 1989). However, they have since been largely replaced by subsequent, PCR-based technologies.

28.4.1.2 Random amplified polymorphic DNA

RAPD are markers based on short, arbitrary primers that amplify genomic DNA if two binding sites are located close to each other (Williams, Kubelik, Livak, Rafalski, & Tingey, 1990). Due to mutations in the primer binding sites, the primers may not bind in some lines, resulting in nonamplification of the DNA. RAPD markers are a dominant marker based on the presence and/or absence of genomic amplicons. The bands are separated by electrophoresis and visualized by DNA staining. RAPD markers are useful for genotyping and fingerprinting of germplasm, because the bands are randomly distributed across the genome (Waugh & Powell, 1992; Welsh & McClelland, 1990).

RAPDs have been successfully used to develop genetic linkage maps and identify markers linked to monogenic traits in breeding programs. For example, RAPD markers were found closely linked to resistance against *Rhynchosporium secalis* in barley (Barua et al., 1993), dwarf and tall coconut palm phenotypes (Rajesh et al., 2013), and resistance against common bean mosaic virus in common beans (Haley, Afanador, & Kelly, 1994). It was noted that as RAPD markers are dominant, it was best to use RAPDs that are absent (in repulsion phase) in the favorable allele, because the absence of a band is then associated with a homozygous allele for the trait of interest, whereas the presence of a marker in coupling phase for the favorable allele could still be heterozygous. However, RAPDs have been shown to have a poor reproducibility and results vary between laboratories (Rajesh et al., 2013; Virk, Ford-Lloyd, Jackson, & Newbury, 1995). Multiple factors can cause this variability in results, such as the short primers and differences in concentrations of primer and template leading to variable primer binding sites.

28.4.1.3 Amplified fragment length polymorphisms

AFLP technology resembles RFLP technology in that genomic DNA is first fragmented by using restriction enzymes (Vos et al., 1995). However, the AFLP protocol differs from RFLP by amplifying and visualizing the fragments by PCR using primers that bind specifically to the restriction site sequences at the ends of the fragments. In this way, generally 50–100 fragments are amplified and separated on polyacrylamide gels. In barley it was shown that the large number of amplified fragments per individual sample resulted in the detection of more genetic diversity by AFLP than by RAPD, SSR, and RFLP marker assays (Russell et al., 1997). AFLP technology is owned by Keygene (Wageningen, NL) and was patented in Europe in 2000 (Vos & Zabeau, 1992).

Similar to RFLP, AFLP is a useful technology for DNA fingerprinting/genotyping of a species of which the genomic sequence is unknown or incomplete. In a study with barley, in which a marker map was developed combining RFLP and AFLP, it was found that AFLP markers seldom interrupted the RFLP clusters but mapped adjacent to them (Becker, Vos, Kuiper, Salamini, & Heun, 1995). Therefore combining RFLP with AFLP markers enriched the marker map of barley.

Furthermore, AFLP is a useful technology to analyze the genetic relationships within a germplasm pool and get marker data to support “essential derivation” of commercial varieties for variety registration (Vuylsteke, Peleman, & Van Eijk, 2007).

28.4.1.4 Simple sequence repeats

SSRs, also called microsatellites, are highly repetitive sequence elements with a short repeat unit (1–8 bp) that vary in length as a result of differences in the number of consecutive repeat units between individuals (Mason, 2015; Tautz, 1989). This variation is thought to be caused by slippage during DNA replication. The variation in microsatellite lengths is used to study genetic lineage and diversity of germplasm and can also be incorporated into genetic maps.

SSRs are useful markers for breeding programs because at every single SSR locus, multiple length-variant alleles can be found, they are evenly distributed across the genome and are codominant. To develop SSR markers, a genomic DNA library first needs to be sequenced and analyzed to detect microsatellites (Edwards, Barker, Daly, Jones, & Karp, 1996). These preparatory steps are time-consuming, but nowadays a lot of sequence information is already available. Using computational tools such as SciRoKo (Kofler, Schlotterer, & Lelley, 2007), SSRs can readily be identified even from partial- or low-coverage genome sequence data (Robinson, Love, Batley, Barker, & Edwards, 2004). Once identified, unique primers are designed based on flanking sequences. SSRs designed in this way have been used to enrich genetic maps of many crop species, and to assess genetic diversity in orphan crops (Ozturk et al., 2017).

28.4.1.5 Sequence characterized amplified region

SCARs are markers that are based on the specific amplification of a DNA region closely linked to a trait of interest, requiring prior knowledge of the DNA sequence of that region. Frequently, RAPD and AFLP markers are converted into SCAR markers by designing unique primers based on the sequence of cloned RAPD or AFLP fragments. Longer primers (15–30 bp) ensure the specificity of the marker locus. SCAR markers are codominant and highly reproducible, which is a major improvement over RAPD markers; SCAR markers are therefore preferred for trait selection/introgression in breeding programs.

For example, two RAPD markers tightly flanking a dominant gene conferring blast resistance in rice, which is caused by the fungal pathogen *Magnaporthe grisea*, were converted into SCARs by designing SCAR primers to the sequenced RAPD amplified product (Naqvi & Chattoo, 1996). Similarly, 8 RAPD markers strongly linked to downy mildew (*Bremia lactucae*) resistance genes in lettuce were successfully converted to SCARs by designing 24-mer primers to the ends of the RAPD products (Paran & Michelmore, 1993). In cowpea, AFLP mapping identified AFLP markers closely linked to the *Rsg1* gene, which confers resistance to *Striga gesnerioides*, a parasitic weed (Boukar, Kong, Singh, Murdock, & Ohm, 2004). A dominant AFLP marker that was in coupling phase with *Rsg1* was converted into a codominant SCAR marker. In tobacco, a similar approach was taken in that AFLP mapping was used to identify amplicons associated with blue-mold (*Peronospora tabacina*) resistance, PVY (Potato Virus Y) susceptibility, and black root rot (*Chalara elegans*) resistance (Julio, Verrier, De, & Borne, 2006), which were then sequenced to design primers and convert them into codominant SCAR markers.

28.4.1.6 Cleaved amplified polymorphic sequences/derived cleaved amplified polymorphic sequences

CAPS and dCAPS are markers that can detect certain SNP. CAPS markers use PCR primers to amplify a specific locus containing an SNP within a restriction site, which is detected by the PCR product being cleaved or not depending on the SNP allele, resulting in different sized bands visible after gel electrophoresis (Baumbusch, Sundal, Hughes, Galau, & Jakobsen, 2001; Konieczny & Ausubel, 1993). The disadvantage of CAPS markers is that they can only detect SNPs that affect a restriction enzyme’s ability to cut DNA (Neff, Neff, Chory, & Pepper, 1998). Therefore an alternative methodology was developed, called dCAPS, which is based on introducing an SNP mutation in one of the primers, which in combination with a specific SNP results in variation in the ability of a restriction enzyme to cut the amplicon.

The design of the primers and restriction enzyme combinations to detect SNP variants is complex. Therefore bioinformaticians have designed online tools to design CAPS/dCAPS markers, such as dCAPS finder 2.0 (Neff, Turk, & Kalishman, 2002), BlastDigester (Ilic, Berleth, & Provart, 2004), and SNP2CAPS (Thiel, Kota, Grosse, Stein, & Graner, 2004).

28.4.2 Shift to single-nucleotide polymorphism and insertion/deletion markers

The markers described earlier were based on restriction digestion and/or PCR amplification followed by separation of the amplicons/fragments by gel electrophoresis. These gels have to be manually assessed for the marker results. Therefore these methods take a lot of manual work to genotype breeding lines and can only be applied to a limited number of lines and for a limited number of markers.

The identification of SNP and short insertions/deletions (INDELs) from sequencing projects opened up the opportunity to develop a new type of markers, which are no longer dependent on time-consuming and costly gel-based assays (Gupta, Roy, & Prasad, 2001). Furthermore, SNPs are the most abundant type of markers, which are widely dispersed across all genomes. Typically, potential SNPs and INDELs are identified from high-throughput sequencing data. Careful bioinformatic analysis is necessary to eliminate sequencing errors that resemble SNPs/INDELs and select those that are widely represented and polymorphic within a population. These markers are useful for all kinds of genotyping, marker-assisted breeding (Kim, Manivannan, Kim, Lee, & Lee, 2019), and fine-mapping for map-based cloning. The high density of SNP markers provides increased mapping resolution which has greatly contributed to map-based gene discovery (Close et al., 2009).

In the last 20 years technologies for screening molecular markers (genotyping) have been significantly improved in their throughput and capacity, especially for SNPs. The development of these higher throughput technologies coincided with an increase in the use of markers in breeding programs. Initially, breeders used markers for single-gene traits, such as single-gene disease resistance. Later, markers were found to be a useful tool to combine multiple single-gene traits into one line, such as multiple disease resistance genes. Here, the codominance of SNP markers became a great advantage, because plants that are homozygous for a trait could be selected in a single generation in a segregating population, saving one generation in time to fix the trait compared with dominant markers such as AFLP and RAPD. Now, breeders are using markers for multiple traits, including single- and complex multigenic traits, for most of their breeding populations. The next step will be full implementation of genomic selection and genomic prediction of phenotypes in breeding programs.

28.4.3 Genotyping technologies and their application in breeding programs

Initially, only large crop breeding programs, such as corn and soybeans, had sufficient funds to adopt the new marker technologies, which they ran in their own labs. More recently, available marker assays have become a lot more affordable and are often provided as a service by specialized companies. This opened the opportunity for breeding programs with smaller budgets and small companies that do not have a molecular lab to start marker-assisted breeding programs. Furthermore, most breeding companies are now using molecular markers for quality control of their varieties or breeds. Parental lines can be verified by unique SNP marker combinations to identify the specific parental germplasm. Seed production departments of breeding companies also check their F₁ hybrid seed for the presence of inbreds (caused by selfings) or contamination with foreign pollen.

In the following sections, the technological advances in genotyping technologies will be reviewed with respect to application in breeding programs, especially stacking of traits, and introgression of traits from wild species using a marker-assisted BC (Section 28.3.1). Attention will be given to the marker efficacy, level of throughput in number of samples and markers, efficiency in the time to deliver the results, and cost efficiency.

28.4.4 Medium-throughput genotyping technologies

The technologies described here all use PCR reactions modified with fluorescent dyes to detect SNPs. Therefore they are accessible to laboratories equipped with real-time PCR machines, or standard PCR machines with fluorescence-capable plate readers. As these assays are usually carried out in a 96-well or 384-well plate format, and fluorescent dyes are relatively expensive, they are most suitable for medium-throughput screening (e.g., a small number of SNPs on several hundred individuals).

28.4.4.1 High-resolution melting

High-resolution melting (HRM) technology is based on amplification in a real-time PCR system of short amplicons (100–200 bp) by specific primers (Simko, 2016; Słomka, Sobalska-Kwapis, Wachulec, Bartosz, & Strapagiel, 2017). The PCR reaction is done in the presence of a dye such as SYBR Green, which has a low level of fluorescence when unbound and high fluorescence when bound to dsDNA. After amplification, melting of the DNA amplicon is observed by measuring the level of fluorescence at incremental temperature increases. Mutations in the amplicon result in its melting at a slightly higher or lower temperature than the reference sequence, giving an altered HRM profile.

HRM is a useful technology for screening populations segregating for known gene alleles, for identifying novel mutations in target genes, and for genotyping of germplasm lines (Simko, 2016). The technology can detect SNPs, INDELs, and SSRs. In particular the ability to detect novel mutations in known target genes is very useful for screening mutant populations and genetically diverse germplasm collections for novel alleles.

In another modification of the HRM protocol, GC tails of different length were added to forward primers specifically binding to a locus containing an SNP to create different melting curves for amplicons from each allele (Wang et al., 2005). This technology achieved a high call rate (98%) and accuracy (99%).

28.4.4.2 *TaqMan—the 5' nuclease assay*

The TaqMan SNP genotyping technology is based on two probes and 2 PCR primers (Francisco, Lazaruk, Rhodes, & Wenz, 2005). The two probes have a higher affinity to bind to the target DNA sequence than the primers and bind specifically to one or the other SNP allele. Each probe is labeled at the 5' end with different fluorescent reporters and at the 3' end with a fluorescence quenching molecule. During the polymerization process, the 5' nuclease activity of Taq polymerase cleaves the 5' reporter dye from the probe, which results in a fluorescent signal. The ratio of one or the other dye indicates the genotype of the tested sample.

28.4.4.3 *Kompetitive allele-specific polymerase chain reaction*

Kompetitive allele-specific PCR (KASP) marker technology can identify SNPs and INDELs and is based on two forward primers that bind specifically to either one or the other allelic variant and have either the HEX (hexachloro-fluoroscen) or FAM (fluorescein amidite) fluorescent molecule attached to them (He, Holme, & Anthony, 2014; Semagn, Babu, Hearne, & Olsen, 2014). The reverse primer is common for both SNP alleles. The reaction is run in a normal thermal cycler, and the fluorescence readings are done on a fluorescence resonance energy transfer-capable plate reader. Plate readouts are analyzed by software that can visualize the marker clusters based on the intensity of the HEX and FAM fluorescent dyes, for example, in three groups for a diploid species: homozygous “AA,” heterozygous “Aa,” and homozygous “aa” or five groups for autotetraploid species: “AAAA,” “AAAa,” “AAaa,” “Aaaa,” and “aaaa.”

KASP is proprietary to Biosearch Technologies (Hoddesdon, United Kingdom) that also produces *SNPline*, a modular high-throughput genotyping system using liquid handling robots and dedicated instruments to fully automate every step of KASP assays, from DNA extraction to SNP detection and genotype scoring.

28.4.4.4 *RNase H2 enzyme-based amplification*

The rhAmp marker assay is very similar to the KASP marker assay, because in both methods two SNP allele-specific forward primers with different fluorescent dyes attached to them bind to either one or the other allele, and a common reverse primer is required to amplify the amplicon (Broccanello et al., 2018; Integrated DNA Technologies I, 2020). The difference with rhAmp is that the primers contain a single RNA residue and a 3' blocking moiety. Only when the primers are perfectly bound to the target DNA, then the RNase H2 enzyme cleaves the blocking moiety from the single RNA base, thereby allowing the polymerase to amplify the PCR product. This blocking moiety increases the accuracy of the assay, because it prevents any primer dimers and off-target amplification, reducing noise in the PCR product.

28.4.4.5 *Accuracy of single-nucleotide polymorphism genotyping*

For all three methods based on differently labeled probes or primers (TaqMan, KASP, rhAmp), it is important that the different marker clusters can be clearly separated. Especially when genotyping tetraploid species, the cluster groups can blend into each other making it difficult accurately to determine the genotype of each individual. In a comparative study, rhAmp and KASP gave a better separation of genotype clusters compared to TaqMan (Ayalew et al., 2019).

In the same study, it was shown that rhAmp had the lowest number of unamplified samples (3%), compared to KASP (6.5%) and TaqMan (7%). In cases of insufficient fluorescence samples get classed as “invalid.” TaqMan had the highest number of “invalid” calls (57 out of 2589) followed by KASP (13 out of 2603) and rhAmp (7 out of 2700). rhAmp had the highest fluorescence signal, which resulted in better separation of the allelic clusters, whereas TaqMan had the least separation of these groups. rhAmp and KASP were also more affordable assays compared to TaqMan (Ayalew et al., 2019; Broccanello et al., 2018). KASP requires more DNA (0.9 ng) in the PCR reaction to get a sufficient fluorescent signal to separate the allelic groups, compared to TaqMan and rhAmp (both 0.2 ng).

28.4.5 High-throughput genotyping technologies

The recent developments on SNP markers have focused on increasing throughput while reducing cost and time through further automation. On the one hand, microarray chips were developed that could run a large number of SNP marker assays simultaneously on one chip. These chips are great for genotyping in detail a relatively small number of samples. This strategy is therefore very suitable for high-resolution genetic mapping and identifying QTLs and markers associated to traits. On the other hand, for breeding applications, there is a clear need for genotyping technologies that can vary the SNPs that are assayed and handle large numbers of samples with a relatively small number of markers. For this purpose, a variety of flexible platforms have been developed that work largely by automating and miniaturizing the PCR-based SNP genotyping reactions described in [Section 28.4.4](#).

28.4.5.1 Diversity arrays technology

In the original DArT protocol, the DNA is cut using a restriction enzymes selected to have target sites in low copy regions, and therefore likely to be close to active genes, and adapters are ligated to these fragment ends ([Jaccoud, Peng, Feinstein, & Kilian, 2001](#)). Amplification is done with primers with selective overhangs, similar to AFLP technology. The amplified fragments are then cloned, further amplified, and added to a DNA microarray. This array could then be used to identify contrasts between two genotype samples by doing the same DNA restriction and PCR steps, but labeling the fragments of one genotype sample with a fluorescent green dye and the other with a fluorescent red dye.

DArT is proprietary technology belonging to Diversity Arrays Technology PL (Australia), which have now switched from using microarrays to screen DArT markers to sequencing-based genotyping, DArTseq ([Edet, Gorafi, Nasuda, & Tsujimoto, 2018](#)). An advantage of both systems is that no prior knowledge of the genomic DNA sequence is required.

28.4.5.2 High-throughput (HTP) fixed single-nucleotide polymorphism microarrays

Fixed SNP microarray chips have the advantage in that they can analyze thousands or even millions of SNP markers for a set of samples at high throughput ([Thompson, 2014](#)). Several companies now provide SNP array technologies.

Illumina's *BeadArray* technology is based on beads coated with specific oligos which are placed in microwells. Initially, SNP detection was done by the GoldenGate assay, which is based on amplification by allele-specific fluorescently labeled primers followed by measuring fluorescence intensity ([Shen et al., 2005](#)). The *BeadArray* chip was replaced by Illumina's higher density *Infinium* chip, which can run 700K SNP assays for 24 samples on one chip and is based on 50-mer oligos and a two-color detection assay ([Steemers & Gunderson, 2007](#)).

Affymetrix offers both conventional SNP microarrays and the high-throughput *Axiom* technology, which similar to the *Infinium* chip combines multiple samples on one chip and is based on a two-color assay of 30-mer probes ([Matsuzaki et al., 2004](#)). *Axiom* arrays have two formats, either genotyping 384 samples with 50K SNPs or 96 samples with 650K SNPs, which equals 4.8 or 62.4 million SNP data points per chip. This is a 25-fold increase in the number of SNP data points compared to single sample SNP chips.

For breeding purposes, the extremely high number of SNPs per sample on microarrays is useful for genome-wide association study (GWAS), diversity analysis, and genomic selection. However, using these array chips in breeding programs is expensive, because of the high design and production cost of the chips. Furthermore, arrays are inflexible once they are designed. If new more relevant SNP marker information becomes available, then a new chip needs to be designed to include these novel SNPs.

28.4.5.3 Fluidigm

Fluidigm's dynamic array, which is marketed under the name Juno, has fully automated the mixing of sample and assay on a nanofluidic platform, which is called an integrated fluidic circuit (IFC). The equipment has also automated the thermal cycling of the SNP assays and harvesting of the product (Fluidigm, San Francisco, United States). The IFC is then scanned on Fluidigm's Biomark or EP1 scanner to collect the SNP marker data. The IFCs come in different sizes of samples \times SNP assays: 48×48 , 96×96 , and 192×24 . KASP, TaqMan, or Fluidigm's own "SNP-type" assays can be run on the dynamic array.

The dynamic array has call rates greater than 99.5% with a call accuracy of 99.8% ([Wang et al., 2009](#)). Each run takes about 3 h to generate 9216 SNP data points (on a 96×96 IFC), which makes it possible to generate 18,432 SNP data points when two runs are done in 1 day. On the IFC, only 6.5 nL reaction volume is mixed into each reaction well of an individual SNP assay. These low reaction volumes make the dynamic array more cost-effective than running the same assays on a PCR machine.

28.4.5.4 Array tape

The principle of array tapes is to decrease reaction volumes to reduce the cost of SNP assays and increase throughput to enhance capacity (Zec et al., 2018). This is achieved by running SNP assays in very small volume droplets on a tape that runs through a fully automated device that dispenses the DNA sample and reaction liquids, runs the thermal cycling, and detects and analyzes the SNP data.

Biosearch Technologies' array tape is a good example of this technology. It is similar to SNPLine, except that reactions are carried out in microwells located on a flexible tape, rather than conventional PCR plates. The tape is handled by a device that is fully automated from setting up the reaction mixes until analysis of the SNP results (LGC_Biosearch_Technologies, 2020). This device can handle 400 arrays of 384 wells per day to deliver 153,600 SNP data points. The number of data points is less than current microarrays, which can handle more than 1 million SNPs, but the system is flexible for which SNP assays are run compared to a fixed microarray. Like Fluidigm, the system also uses only small reagent volumes, as 800 μ L is required to run 200 plates of 384 wells, which is equivalent to about 10 nL per individual SNP assay.

28.4.5.5 OpenArray

Applied Biosystems' OpenArray is a semiflexible microwell array, because the customer needs to order the SNP assays of their choice in advance (Broccanello, Gerace, & Stevanato, 2020; Thompson, 2014). The microplates are then prepared by the supplier. Each array can run 3072 TaqMan SNP assays with a volume of 33 nL per reaction, with a capacity to produce 70,000 SNP data points per day. The OpenArray DLP Real-Time qPCR platform gives the additional advantage to quantify the DNA, which is useful for applications such as pathogen detection and quantification (van Doorn et al., 2007).

28.4.5.6 iPLEX Gold assay

The iPLEX Gold assay, which is provided by the US company Sequenom, is a high-throughput marker assay technology based on a single-base extension from a specific primer pair (Gabriel, Ziaugra, & Tabbaa, 2009; Perkel, 2008). Depending on the SNP, the sequence is extended by four terminator bases that differ in molecular weight (12 Da difference in weight). A mass spectrometer is used to detect the differences in molecular weight using two 384-position matrix-assisted laser desorption/ionization target plates. In total, 10 plates can be processed per day, which equals more than 138,000 SNPs.

The advantage of this technology is high-throughput and automation. Also, it is highly sensitive and can detect very small quantities of DNA, as demonstrated on ancient human DNA samples (Mendisco et al., 2011).

28.4.5.7 Genotyping-by-sequencing

While all the methods described earlier require advance knowledge of SNP-containing sequences, genotyping-by-sequencing (GBS) exploits high-throughput DNA sequencing platforms both to screen known SNPs and discover new ones. GBS is a reduced-representation sequencing method (Elshire et al., 2011) that streamlines the construction of libraries from an earlier GBS-type methodology, named Restriction Association DNA sequencing (RADseq) (Baird et al., 2008). GBS has gained popularity due to the continuously reducing cost of sequencing, especially with the Illumina next-generation sequencing platform. Since then it has developed into a useful methodology that is able to discover a large number of genome-wide SNPs, and genotype germplasm (Peterson, Dong, Horbach, & Fu, 2014). GBS can be applied without prior genome sequence information and is therefore useful for crops and plant species with no genomic sequence data.

The GBS methodology is based on creating a reduced representation of the genome by cutting the genomic DNA with restriction enzymes (Peterson et al., 2014; Poland & Rife, 2012). Enzyme-specific adapters are then ligated to the fragments, followed by a PCR amplification of the fragments, which are barcoded per sample to enable multiplexing and pooled into a library. The library is sequenced, and the sequence reads are assembled and aligned with each other and/or a reference genome. Software packages such as Samtools (Li et al., 2009) are used to identify variant locations from the alignments, which are stored in Variant Call Format (VCF) files. The VCF files are then mined for SNPs.

The whole process takes up to 2 months to complete the discovery of novel SNPs and genotype 300 samples, with a cost of \$20,000–\$30,000 for the sequencing (Peterson et al., 2014). GBS is a good methodology to generate large numbers of SNPs, which enables QTL mapping and GWAS to identify markers closely associated with a trait locus

(Siddique et al., 2019), or even gene discovery through high-density marker maps (Paulsmeyer, Brown, & Juvik, 2018), although the other methods described earlier remain more cost-effective for routine screening of known SNPs.

28.4.6 Increased automation and throughput while reducing cost per data point

Genotyping assays such as RFLP, RAPD, AFLP, SSR, and SCAR were based on marker types that were present at relative low-frequency throughout genomes. Furthermore, these early marker assays were laborious, and therefore more costly per sample, as the detection was based on DNA separation on gels (Table 28.2). They are not readily automated, although pipetting robots can be used for some steps.

The discovery of SNP markers was a major breakthrough for genetic research and breeding, because SNPs are very frequent, codominant by nature, and evenly dispersed throughout genomes allowing for the identification of markers closely linked to traits of interest. SNP marker assays such as CAPS/dCAPS were still based on DNA separation on gels and could only use SNPs located in restriction sites. The development of protocols using fluorophore-tagged PCR probes and primers made it possible to detect SNPs by contrasting fluorescence intensities. This could now be applied in 96-well up to 1536-well plate formats to speed up SNP detection. Since then, the SNP assay capacity was further increased either through DNA microarrays with capacity to detect very large numbers of SNPs (up to 2.6 million on the Axiom platform) but only on a relatively small number of samples, or through the development of flexible microwell arrays where the user is able to select different SNP \times sample combinations in each run and that can handle larger numbers of samples simultaneously, which is more suitable for breeding programs. These technologies are summarized and compared in Table 28.3. High-throughput SNP technologies greatly reduce the cost of genotyping per data point; however, they do require substantial investment in specialist equipment, and/or to produce the individual SNP assay reagents.

With the reducing cost of sequencing, it is now much more affordable to sequence a reduced representation of the genome, which is what has been achieved both by the GBS protocol and by DArTseq. Both discover novel SNPs and genotype samples in the same experiment, with no prior genome sequence data required. However, they do require additional bioinformatic expertise in order to correctly identify SNPs within the sequence data.

28.4.7 Single-nucleotide polymorphism genotyping for sustainable agriculture in a complex genome—bread wheat

Bread wheat (*Triticum aestivum* L.) is one of the world's most important crops in terms of calories contributed to the human diet. However, it is a challenging crop for both researchers and breeders owing to its large (1C = 16 Gb), hexaploid, and highly repeat-rich genome. In the light of this, considerable efforts have been invested over the last 15 years, most significantly in the framework of the IWGSC (International Wheat Genome Sequencing Consortium, <http://www.wheatgenome.org>) to produce a high-quality reference genome assembly for the cultivar “Chinese Spring” (Appels et al., 2018). These efforts have culminated in the IWGSC RefSeq v2.0, which is currently available for download under a Data Access Agreement. The IWGSC RefSeq v1.0 and annotation v1.1 are already available for open access through URGI (<https://wheat-urgi.versailles.inra.fr/>) along with marker and phenomic data (Alaux et al., 2018). An important challenge for bioinformaticians is to mine this large amount of available information for specific data that is valuable for wheat breeders, such as molecular markers. As described above, SNPs are arguably the most readily validated markers in many crops. One of the major environmental problems affecting the sustainability of wheat production is drought, therefore SNPs associated with drought tolerance would be of great value.

Although considerable research has been directed toward drought stress in wheat, identifying informative loci is not trivial. For example, a search of the NCBI databases using the terms “drought” and “Triticum[Organism]” (<https://www.ncbi.nlm.nih.gov/search/all/>, searched on 04.12.2020) identifies 518 SRA datasets (predominantly RNA-seq studies in *T. aestivum*), over 117,700 nucleotide records (mostly mRNA transcripts assembled from the abovementioned experiments) but only 21 genes that are annotated as having a role in drought stress. This reflects the complexity of the drought stress response, where many genetic loci are involved but the contribution of each varies considerably depending on the environmental conditions and genetic background (Tardieu, 2012).

The first draft assembly of the bread wheat genome, generated from the cultivar “Chinese Spring” using 454 sequencing, was released as long ago as 2012 (Brenchley et al., 2012). Although highly fragmented and at low average sequence coverage ($5\times$), this allowed assembly of many protein-coding genes and identification of SNPs distinguishing between the three subgenomes of hexaploid wheat (A, B, and D). Subsequently, improved assemblies were constructed from specific sequencing of individually flow-sorted chromosomes (Lucas et al., 2012; Mayer et al., 2014), and

TABLE 28.2 Summary of low-throughput genetic markers.

Marker type	Assay type	Genomic sequence information required?	Dominant/codominant	Detection method	Sample × marker capacity	Reaction volume (=cost) (μL)	Remarks	References
Restriction fragment length polymorphism (RFLP)	Restriction enzyme and probe	No	Codominant	Gel electrophoresis and radioactive probe	Limited by lanes on gel	50		Beckmann and Soller (1986), Haanstra et al. (1999)
Random amplified polymorphic DNA (RAPD)	Polymerase chain reaction (PCR)	No	Dominant	Gel electrophoresis and DNA binding dye	Limited by PCR machine capacity and/or lanes on gel	25	Marker assay is highly sensitive to changes in experimental conditions	Virk et al. (1995), Eujayl, Baum, Powell, Erskine, and Pehu (1998), Joobeur, Periam, de Vicente, King, and Arús (2000)
Amplified fragment length polymorphisms (AFLP)	Restriction enzyme and PCR	No	Dominant/Codominant ^a	Gel electrophoresis and silver staining or radioactive labeling		11	Most bands are monoallelic (absence/presence). Biallelic bands occur at low frequency	Vuylsteke et al. (2007), Vuylsteke et al. (1999)
Simple sequence repeat (SSR)	PCR	Yes	Codominant	Gel electrophoresis and DNA binding dye		25	Multiallelic at a single locus due to SSR amplicon length variation	Mason (2015), Robinson et al. (2004), Varshney et al. (2007), Hwang et al. (2009)
Sequence characterize amplified region (SCAR)	PCR	Yes	Codominant	Gel electrophoresis and radioactive labeling		10	RAPD and AFLP markers are often converted into SCAR markers for increased reproducibility and codominance	Paran and Michelmore (1993)
Cleaved amplified polymorphic sequence (CAPS)/dCAPS	PCR and restriction enzyme	Yes	Codominant	Gel electrophoresis and DNA binding dye		30	Optimization of PCR amplification required for each individual CAPS marker	Baumbusch et al. (2001), Neff et al. (1998)

^aAFLP bands can be scored codominantly based on intensity of the bands. However, there is a risk of wrong interpretation.

TABLE 28.3 Summary of medium- and high-throughput genotyping technologies.

Technology	Marker types	Detection method	Automation level	Sample × marker capacity	Reaction volume (=cost)	Cost	Remarks	References
<i>HRM</i> —high-resolution melting	Single-nucleotide polymorphism (SNP), INDEL, simple sequence repeat (SSR)	Real-time polymerase chain reaction (PCR) and fluorescence	Pipetting robot could be used. Thermal cycling and fluorescence detection automated.	Limited by real-time PCR capacity	20 µL	\$0.012—\$0.014 per data point	AA, Aa, aa genotypes generally have different melting curves.	Simko (2016) , Słomka et al. (2017) , Galuszynski and Potts (2020)
<i>TaqMan</i> —the 5′ nuclease assay	SNP, less than 6 bp INDEL	Real-time PCR and fluorescence			5 µL	\$0.41 per SNP data point (2018 data)		Francisco et al. (2005) , Ayalew et al. (2019)
<i>KASP</i> —kompetitive allele-specific PCR	SNP, INDEL	PCR and fluorescence	Option to use LGC’s SNPLine, which automates all steps from DNA extraction to marker detection.	Limited by PCR capacity. SNPLine’s capacity: 145,000 data points per day, ideal for many samples and few SNPs	5 µL	\$0.15 per SNP data point (2018 data)	Even though KASP is run as an uniplex; it can be run across multiple plates (from 96- to 1536-well) with different samples/SNP primer combinations.	Semagn et al. (2014) , Ayalew et al. (2019)
<i>rhAmp</i> —RNase H2 enzyme-based amplification	SNP, INDEL	Real-time PCR and fluorescence	Pipetting robot could be used. thermal cycling and fluorescence detection automated.	Limited by real-time PCR capacity	5 µL	\$0.12 per SNP data point (2018 data)	<i>rhAmp</i> gave a better separation of the homozygous and heterozygous genotype clusters compared to KASP and <i>TaqMan</i> .	Broccanello et al. (2018) , Ayalew et al. (2019)
<i>DArT</i> —diversity arrays technology	Polymorphic restriction fragments, some SNPs	<i>DArTarray</i> : Fluorescence-labeled DNA microarray <i>DArTseq</i> : Sequencing	Automated fluorescence detection/sequencing	Can detect variation for a large number of DNA fragments for a relatively small number of samples.	5 µL	\$0.10 per data point		Jaccoud et al. (2001)

(Continued)

TABLE 28.3 (Continued)

Technology	Marker types	Detection method	Automation level	Sample × marker capacity	Reaction volume (=cost)	Cost	Remarks	References
<i>BeadArray</i>	SNP	Microarray and fluorescence detection	Automated fluorescence detection and scoring of genotype.	In 1 day: 2 array matrices with 96 sample × 1536 SNP loci = 295,000 data points	17 nL			Shen et al. (2005), Steemers and Gunderson (2007)
<i>Infinium chip</i>	SNP	Microarray and fluorescence detection	Automated robotic pipetting, array hybridization and fluorescence detection	700,000 SNP loci × 24 samples per chip	1.1 nL	\$0.001 per SNP data point		Steemers and Gunderson (2007)
<i>Axiom</i>	SNP	Microarray and fluorescence detection		1500–2.6 million SNP markers per array with 24–96 arrays per plate	0.1 nL	\$0.000025 per SNP data point		Matsuzaki et al. (2004)
<i>Fluidigm</i>	Flexible SNP assay	Two-color fluorescence detection	Automated robotic pipetting and sample and reaction liquid mixing on the IFC	Several different IFC's, but largest can handle 96 samples × 96 markers = 9216 data points	6.5 nL			Yu et al. (2020)
<i>Array Tape</i>	Flexible SNP assay	Fluorescence detection	All processes from mixing the samples and reaction liquids, thermal cycling to detection of the SNP's is fully automated	Totally flexible on combinations of sample × SNP markers. In 1 day this equipment can run 400 arrays of 384 wells = 153,600 SNP data points	10 nL			Zec et al. (2018)
<i>OpenArray</i>	Flexible SNP assay	Fluorescence detection	Automated robotic pipetting; in each of the 48 subarrays the sample is automatically divided over 64 individual SNP assays through microholes.	48 samples × 64 SNP markers = 3072 SNP data points	33 nL		Dependency on using TaqMan assays.	Broccanello, Gerace, and Stevanato (2020), van Doorn et al. (2007)

<i>iPLEX Gold</i>	Flexible SNP assay	Mass spectrometer	The thermal amplification and SNP detection in the mass-spec is one fully automated workflow.	Flexible on sample × marker combinations; Capacity of 138,000 SNP's per day	0.07 nL			Perkel (2008) , Gabriel et al. (2009)
<i>GBS—genotyping-by-sequencing</i>	All sequence polymorphisms	DNA sequencing	Pipetting robot could be used. Sequencing reactions and data collection are automated.	Limited by sequencing library cost and platform capacity, it can theoretically identify all SNPs present in sample	NA	Per-sample cost of sequence library preparation is limiting factor	GBS discovers new SNPs and genotypes the samples simultaneously. Bioinformatics support is important to analyze the sequence data.	Peterson et al. (2014) , Poland and Rife (2012)

new SNPs were identified through resequencing of multiple bread wheat varieties (Allen et al., 2012) and its wild relatives (Akpınar, Lucas, Vrána, Doležel, & Budak, 2015). The fragmentary nature of the draft genomes meant that comparative genetic mapping from fully sequenced model grass species such as *Brachypodium distachyon* was essential to determine the gene order (Lucas et al., 2013) and new pipelines were developed to identify SNPs from alleles that could vary from 1 to 6 copies between individuals (Akpınar, Lucas, & Budak, 2017). Over 850,000 SNPs discovered in these studies were then used to construct a large-scale Axiom genotyping array (Winfield et al., 2016). The SNPs that were observed to be most informative (present and polymorphic in a wide variety of germplasm) were then incorporated into a smaller “Wheat Breeders” microarray (35,000 SNPs) that is more efficient for high-throughput genotyping (Allen et al., 2017). This array has been validated using over 6000 cultivated wheat varieties and wild relatives. Its utility for trait mapping has been demonstrated both in segregating bread wheat populations (Hussain, Lucas, Ozturk, & Budak, 2017) and in an exotic population of RILs produced by crossing tetraploid Durum wheat cultivars with wild relatives (Lucas, Salantur, Yazar, & Budak, 2017). In these studies, once the SNP genotyping analysis pipelines were adapted to the germplasm being used, genetic maps were constructed and QTLs mapped for complex traits such as tolerance to salt stress, osmotic stress, yield, and antioxidant production.

In summary, the last 10 years of genome research in bread wheat has generated an enormous quantity of sequencing data from which many different genetic markers and high-throughput genotyping tools have been developed. Bioinformatics played an essential role at every step of this process, mining and integrating many different kinds of data in order to determine functional differences that can be used to improve sustainability of wheat production, for example, through MAS. Furthermore, this case study illustrates that even when genomic data is limited or incomplete—as in many other agriculturally important species—valuable biological information can be obtained through methods such as comparative mapping. However, much of this data remains in the academic domain; there remains a need for computational analyses that streamline the translation of these findings into tools that can be applied in the field.

28.5 Genome-wide association studies

The causal correlation between genetic polymorphism and the phenotypical variations found between individuals is of basic biological importance. Advancing knowledge of both the specific loci underlying a phenotype and the genetic design of a function increases the capacity to predict genetic factors underlying essential agronomic traits such as growth rate and yield for plants. While genetic mapping (Section 28.3) aims to uncover such genotype–phenotype relationships in a carefully constructed mapping population with known heredity, association mapping explores the diversity created in nature by hundreds of generations of natural selection, or obtained through random mutagenesis. The concept is to find statistical correlations (association) between genetic changes and phenotypes that reoccur in diverse, distantly related individuals. Using an appropriate statistical framework, phenotypes are linked back to the underlying genetic loci to define QTLs. From this viewpoint, GWAS is considered an effective and complementary method to genetic mapping, to connect the genotype–phenotype map (Korte & Farlow, 2013).

28.5.1 Using single-nucleotide polymorphism markers for genome-wide association studies

The fundamental approach in GWAS is to estimate the association between each genotyped marker and phenotype(s) that have been identified among multiple individuals. Molecular markers are reliable tools for detecting population structure in a collection of genotypes. However, the ability to detect population structure and genotype–phenotype associations depends on using a large number of molecular markers distributed throughout the whole genome. Therefore SNP markers are widely used due to the advent of HTP genotyping methods (described in Section 28.3) in both plant and animal genomes (Hiremath et al., 2012; Rimbart et al., 2018). SNPs are advantageous for GWAS due to their wide distribution in the genome, codominant transmission, chromosome-specific location, and high reproducibility (Kujur et al., 2015). Therefore the initial phase of GWAS often includes SNP discovery and capture using HTP sequencing, such as the GBS approach described in Section 28.4.5.7 (Elbasyoni et al., 2018), or similar methods such as DArTseq or double-digest RADseq (Helmstetter, Oztolan-Erol, Lucas, & Buggs, 2020). All of these GBS-like methods are based on “reduced-representation” sequencing, in which the genome is cut with restriction enzymes and sequencing is performed starting from chosen restriction sites (Thompson, 2014). Restriction sites are relatively well dispersed throughout the genome, so this provides a random sampling of tens of thousands of sequence tags. Not all restriction sites are conserved between any pair of individuals, so in principle it is also possible to score the presence/absence of specific restriction sites, similar to RFLP/AFLP markers. For GWAS however, it is usually SNPs that are correlated with phenotypes or traits based on the GWAS design.

GWAS is usually carried out on a “diversity panel” consisting of germplasm selected to represent as much of the genetic variation present in a natural population as possible. Therefore SNPs in the panel are mutations that may have occurred thousands of generations ago and spread by natural selection or chance. When a second SNP is formed close enough to a previously existing SNP that recombination between them is rare (which may be more than tens of thousands of base pairs), these two variant alleles are often passed on to individuals in the next generation. This noncoincidental association of two alleles is defined as linkage disequilibrium (LD). If the presence of a specific SNP is genetically linked to, for example, increased susceptibility to a disease, a statistically significant association between the disease and that SNP (directly related) and several nearby SNPs (indirectly related by LD) is observed. Therefore each individual in the diversity panel is both genotyped and phenotyped for the trait(s) of interest, and statistical correlations between SNPs and phenotypes calculated within the panel, using single-locus or multilocus tests (Bush & Moore, 2012). GWAS results are typically obtained in a shorter time compared to traditional genetic mapping methods, as the study is carried out on an already existing natural population. However, false-positive results can arise due to unknown population structure and kinship within the diversity panel (Sun et al., 2016; Turner et al., 2017). Therefore it is of great importance to develop statistical methods and bioinformatic programs that effectively model the population structure and relationships between individuals in association mapping studies (Mangini et al., 2018; Zhou et al., 2017).

28.5.2 Genome-wide association studies’ design and analysis

GWAS experiments are genetic mapping studies used to identify markers associated with desired traits based on the principle of LD between allelic variants in genetically highly diverse natural populations and the trait of interest (Mangini et al., 2018; Warmerdam et al., 2018). Compared to these populations, biparental genetic linkage mapping populations (see Section 2.2) have lower resolution due to the relatively small number of generations for recombination and segregation (Collard, Jahufer, Brouwer, & Pang, 2005; Nadeem et al., 2018). Association mapping populations are not required to be created, saving time compared to traditional mapping methods (Turner et al., 2017). The successful application of GWAS relies on study design, genotyping technologies and statistical concepts for analysis, copying, interpretation, and tracking of association results (Bush & Moore, 2012).

A typical genome-wide association study consists of four stages. These are the selection of the population, DNA isolation and genotyping, statistical evaluation of the relationship between the phenotypes and SNPs that exceed a threshold value, and finally, the repetition of the defined relationship in independent population samples to verify the association or to examine the function experimentally. Standard methods or designs are required for an efficient GWAS. In crops, GWAS often uses a diversity panel as described above, whereas in livestock a case–control design may be more appropriate, in which cases are ascertained based on the trait of interest (Cardon & Bell, 2001). The case is the group exhibiting a desired trait, while the control group does not include the trait from the case group. The advantages of this design are that cases are readily obtained and can be efficiently genotyped and compared with control populations. The selection of controls is a critical step because any systematic difference in allele frequency between cases and controls can only be interpreted in relation to the trait of interest, but it could also result from independent processes such as evolutionary background, sex differences, and domestication practices (Cardon & Bell, 2001). Choosing sample numbers and selection of samples for each group is very important for GWAS analysis. The number of samples varies depending on the type of study, the characteristics of the groups, and factors such as location. Also, samples in the same group must be homogeneous in terms of the specified traits.

Population structure is the result of selection and mixing in a population, leading to a high level of observed LD between physically unlinked markers (Rostoks et al., 2006) and can be used to infer relationships between individuals within a population and between different populations. It also gives an insight into the evolutionary relationships of individuals in a population. Various approaches have been proposed in association mapping studies to estimate a population structure, including distance-based and model-based methods. Generally, genetic analysis is a measure of population substructure (Bush & Moore, 2012). Population structures within the GWAS population are analyzed to avoid stratification and are subjected to principal component analysis to minimize effects within the data. Also, the STRUCTURE program was developed to determine the population structure and to subgroup individuals using the Bayesian approach (Pritchard, Stephens, & Donnelly, 2000).

When quantitative traits are assessed in GWAS analysis, logistic regression can be used for the analysis of characteristics. Programs such as TASSEL can be used to establish relationships between quantitative phenotypic characters and genotype (Bradbury et al., 2007; Thornsberry et al., 2001). TASSEL incorporates tools to normalize the results according to known or inferred population structure, and perform analyses according to generalized linear model (GLM) and mixed linear model (MLM) models. In a case–control GWAS design for a binary trait, a feature analysis using the GLM is carried out and

analysis of variance is performed to find statistically significant associations (Bush & Moore, 2012). For a study of multiple quantitative traits in a diversity panel, MLM is more appropriate (Lucas et al., 2017).

One application of modern phenotyping technologies to sustainable crop development is to combine image-based “phenomics” with GWAS. In this approach, photographs of the plants are taken at regular intervals in a phenotyping platform, and image-based traits (i-traits) extracted by automated image analysis. GWAS analysis combined with i-traits obtained from images enables physiological and morphological traits to be phenotyped more frequently without harming or perturbing the plants, and some i-traits were shown to be well correlated with field drought tolerance in rice (Guo et al., 2018).

28.5.3 Applications of genome-wide association studies to plant and animal breeding

Since the first association mapping study reported for an agriculture species, carried out in maize (Bar-Hen, Charcosset, Bourgoin, & Guiard, 1995), GWAS experiments have been carried out on many plants, particularly to evaluate agronomic characters that depend on many genes (Xiao et al., 2018). GWASs are now routinely applied in a range of model organisms including *Arabidopsis* (Atwell et al., 2010) and mouse (Flint & Eskin, 2012), and to nonmodel systems including crops (Huang et al., 2012; Ranc et al., 2012; Wang et al., 2012) and cattle (Olsen et al., 2011). Table 28.4 summarizes some of the most recent applications of GWAS to study agronomically important traits for both plants and animals.

TABLE 28.4 Examples of GWAS conducted on different plant and animal species in recent years.

	Species	Target traits	Outcomes	References
Plant	Mung bean	Salt stress	A total of 5288 single-nucleotide polymorphism (SNP) markers obtained through genotyping-by-sequencing were used to mine alleles associated with salt stress tolerance.	Breria et al. (2020)
	Rice	Cold tolerance	Five genetic loci at the booting stage and eight genetic loci at the seedling stage related to cold tolerance have been identified.	Xiao et al. (2018)
		Drought tolerance	Identified 69 image-based trait (i-traits) locus associations by both genome-wide association studies (GWAS) and linkage analysis of a recombinant inbred line population. The relevance of i-traits to drought tolerance in the field.	Guo et al. (2018)
		Salinity tolerance	At the reproductive stage under salt stress, SNPs associated with a number of phenotypic traits and several related genetic loci were identified. 1200 candidate genes have been identified from many functional categories, including cation transporters and transcription factors with a known role in salinity tolerance.	Kumar et al. (2015), Patishtan, Hartley, Fonseca de Carvalho, and Maathuis (2018)
		Resistance to rice blast (<i>Magnaporthe grisea</i>)	Using both field and growth chamber screenings, reported 14 marker–trait associations for blast resistance. A total of 11 accessions were discovered showing high resistance in both field and controlled conditions.	Volante et al. (2020)
	Alfalfa	Salt tolerance	Identified 27 SNP markers associated with salt tolerance. Mapping of SNP markers against the reference <i>Medicago truncatula</i> genome revealed multiple putative candidate genes based on their functional annotations.	Medina, Hawkins, Liu, Peel, and Yu (2020)
	Barley and	Yield traits	A quantitative trait loci on chr.4H associated with powdery mildew and <i>Ramularia</i> resistance was	Tsai et al. (2020)

(Continued)

TABLE 28.4 (Continued)

	Species	Target traits	Outcomes	References
	winter wheat		identified in spring barley. For winter wheat, two SNPs on chr.6A and one SNP on chr.1B were significantly associated with moisture quality trait, while one SNP on chr.5B associated with starch content in seeds.	
	Cotton	Drought resistance	Determined 390 genetic loci by GWAS using 56 morphological and 63 texture i-traits.	Li et al. (2020)
Animal	Cattle dog	Deafness	One important genome-wide association and 14 indicative (chromosome-wide) associations of bilaterally deaf Australian cattle dogs were reported using GWAS.	Hayward et al. (2020)
	Pig	Fatty acid composition Meat and carcass traits	Information was provided on mutual genetic regulation of the composition of fatty acids and other economic characteristics in pork. The variant effect estimation revealed that 15 high effect variants for back fat thickness, meat–fat ratio, and carcass length traits were among the statistically significant associated variants.	Zhang et al. (2019) , Falker-Gieske et al. (2019)
	Cattle	Disease traits	Reported six significant associations and 20 candidate genes of cattle health	Freebern et al. (2020)
	Buffalo	Milk yield	Four significant regions were identified associated with milk yield.	El-Halawany et al. (2017)
	Sheep	Coat color	The genomic region, candidate gene, and genetic variants associated with the coat color phenotype in Chinese Tan Sheep were identified an ovine 600K SNP BeadChip.	Gebreselassie et al. (2020)

28.6 Emerging strategies for breeding and genetics

Bioinformatics emerged as a tool to facilitate biological discoveries more than a decade ago. With the advancement of the Human Genome Project, the ability to relate sequencing data with biological process evidence has improved enormously, but many gaps in our knowledge still remain. It is therefore more necessary than ever to be able to collect, manage, store, evaluate, and interpret data. Although a relatively new field of study, bioinformatics is advancing as quickly as any sector of biotechnology ([Xue, Zhao, Liang, Hou, & Wang, 2008](#)).

In human genetics, bioinformatics is widely used in medicine to determine the genetic details of different diseases. The field of agriculture has also taken advantage of these studies as microorganisms play a significant role in agriculture, and bioinformatics can analyze complete genome information of these species. Agriculture has since gained from the sequencing of many genomes of plants and livestock. Bioinformatics techniques play an essential role in converting raw sequence data into structures of the genes found in these organisms. These tools often allow the functions of various genes and factors influencing these genes to be predicted. The knowledge generated about these the genes helps scientists develop improved crop species, for example, with drought, herbicide, and pesticide tolerance ([Bhattacharyya, Goswami, & Bhattacharyya, 2016](#)). Similarly, in livestock, specific genes may be mutated to increase meat and milk development. Certain genome modifications could also be introduced to make them immune to disease ([Murray & Anderson, 2000](#)). In this section, we summarize emerging approaches to improving agriculturally important traits that are facilitated by bioinformatics.

28.6.1 Gene expression regulation by noncoding RNA

Although a large proportion of DNA, which is the genetic material in living things, can be transcribed into RNA, a very small amount (approximately 1.5% depending on species) of the genome is used in functional protein synthesis.

Parts that are not translated into protein and expressed as noncoding RNA (ncRNA/noncoding RNA) have until recently been considered of little importance (Wijnhoven, Michael, & Watson, 2007) and largely ignored in many cases. However, in recent studies, it has been revealed that there are many types of ncRNAs that play a role in important biological events such as the regulation of gene expression (Morris & Mattick, 2014). There are three types of ncRNAs that can activate RNAi: endogenous micro-RNA (miRNA/miR), exogenous small interfering RNA (siRNA), and piwi-interacting RNA (piRNA) in germ cells. Although plants do not contain piRNAs, the third largest class of small RNA found in animals, they have expanded their repertoire of endogenous siRNAs, some of which fulfill similar molecular and developmental functions to piRNAs in animals (Chen, 2012; Ku & Lin, 2014). These RNAs have important roles in posttranscriptional gene regulation. Messenger RNAs, which are the direct products of protein-coding genes, can reduce (by inhibition of translation) or occasionally increase their activity on binding to these RNAs. RNAi-inducing ncRNAs are encoded by genes that are transcribed from DNA but not translated into protein (Gupta, 2014).

The most well-studied of these RNAi molecules are miRNAs, which are versatile regulators of gene expression in both plants and animals (Yang, Farmer, Agyekum, & Hirschi, 2015). Starting with the transcription of DNA in the nucleus via RNA polymerase II enzyme, miRNAs are 21–24 nucleotide-long ncRNAs that are processed from a longer mRNA-like primary transcript (pri-miRNA) via a hairpin-structured intermediate (pre-miRNA) (Cech & Steitz, 2014; Morris & Mattick, 2014). The first study investigating miRNAs was conducted by Lee, Feinbaum, and Ambros (1993) on *Caenorhabditis elegans*, a round worm. This nematode was screened for gene content and it was observed that a gene named *Lin-4* does not encode any protein but reduced the expression of a target gene. With the discovery of miRNAs in other organisms in the early 2000s, it was understood that the miRNA-based transcriptional regulation mechanism has a general and important role in the developmental process and it has become a focus of attention for scientists, industry, and the private sector in recent years. As more sequencing data have become available, the number of newly discovered miRNAs has increased significantly (Kozomara, Birgaoanu, & Griffiths-Jones, 2019), with the aid of bioinformatic tools to predict miRNAs and their targets in both plants and animals (Dai, Zhuang, & Zhao, 2018; Lucas & Budak, 2012).

It is known that miRNAs control the expression levels of genes related to important sustainability traits such as growth and stress tolerance in plants, acting as endogenous gene regulators (Han et al., 2014). This regulation is implemented through posttranscriptional degradation, translation inhibition (Rogers & Chen, 2013), methylation (Chellappan et al., 2010; Wu et al., 2010), or histone modification (Chuang & Jones, 2007). In addition to regulating endogenous genes, miRNA and other small RNAs also help preserve genome integrity by suppressing genetic material such as transposons, retrotransposons, and viruses (Tomari et al., 2004). It is known that miRNAs play an active role in diverse biological processes, including organ development (Aukerman & Sakai, 2003), hormone signaling (Mallory, Bartel, & Bartel, 2005), defense against pathogens (Navarro et al., 2006), and response to abiotic stress (Sunkar, Chinnusamy, Zhu, & Zhu, 2007). These abiotic stresses include salinity (Covarrubias & Reyes, 2010), drought (Eldem et al., 2012; Li, Qin, Duan, Yin, & Xia, 2011), cold (Zhou, Wang, Sutoh, Zhu, & Zhang, 2008), and heavy metals (Huang, Peng, Qiu, & Yang, 2009). For example, it has been observed that the expression of zma-miR169 in maize decreases during drought (with ABA and PEG application) and increases first and then decreases during salinity stress (NaCl application) (Luan et al., 2015). In rice plants, miR319 has been shown to play a role in cold tolerance by targeting the TEOSINTE BRANCHED/CYCLOIDEA/PCF (TCP) transcription factor (Yang et al., 2013). In a study conducted on tomato plants, it was found that the expression of miRNAs targeting genes that provide abiotic stress adaptation and disease resistance decreased following treatment with abscisic acid (ABA), a signal molecule that activates the stress response (Cheng et al., 2016). These results show that the miRNA-mediated stress responses can be regulated by external ABA treatment. All such studies show that miRNA-based RNAi technology is an effective tool in many areas targeted by plant biotechnology. The future development of genomic tools and techniques for the detection of new miRNAs in different agricultural species will help us to better understand miRNA-mediated gene regulation during diverse abiotic stresses (Shafi & Zahoor, 2019).

Given their importance, it is not surprising that multiple bioinformatic databases and tools have been developed focusing on the functions of miRNAs in crop plants, including PASmiR, a complete repository of miRNA pathways for controlling abiotic stress reactions in the plant stress physiology community (Zhang et al., 2013); PmiRExAt, an online database resource that provides a plant miRNA atlas; and WMP, a resource that offers data on abiotic stress-responsive miRNAs in wheat (Remita et al., 2016). Furthermore, the high conservation of miRNAs between related species means that cross-species studies are possible, for example, a miRNA microarray developed for barley was used to identify drought-related miRNAs in wild emmer wheat (Kantar, Lucas, & Budak, 2011).

In animals, miRNAs play a role in many important cellular processes such as cell proliferation, differentiation, and apoptosis (Liu, Song, Chen, & Yu, 2009; Peng, Zhao, Shen, Mao, & Xu, 2015), and most miRNA research has focused on elucidating the mechanisms of cancer in humans. In animals, a total of 4312 experimentally validated miRNAs from

different species have been identified so far (Huang et al., 2020). There are also miRNA profiling studies in economically important livestock, and it has been reported that miRNAs are effective in tissue and organ development, shaping the immune response and metabolic events. In addition, there are studies showing that SNPs found within miRNA genes are associated with phenotypic differences between animals, yield traits, and susceptibility to diseases (Jevsinek Skok et al., 2013). For example, studies on miRNAs in cattle focus on adipose tissue, skeletal muscle, oocyte development, early embryonic development, milk yield, and mastitis (Wenguang, Jianghong, Jinquan, & Yashizawa, 2007). By comparing the miRNA profiles of different stages of lactation in goats, miRNA identification was performed in mammary gland, milk, and colostrum (Hou et al., 2017; Mobuchon et al., 2015; Shao et al., 2014). There are also miRNA studies on bone and cartilage development and glycogen metabolism in horses (Desjardin et al., 2014; Gim et al., 2014). Over 800 miRNAs have been identified in studies in chickens, and some were reported to influence embryonic development, skeletal muscle, and adipose tissue development (Xu, Wang, Du, & Li, 2006). In pigs, miRNAs related to the reproductive system, skeletal muscle, fat development, and immune system have been identified (Wenguang et al., 2007). Many economically important yield characteristics in farm animals are under the influence of many genes and are quantitative. Studies on miRNAs in cattle, sheep, goats, horses, chickens, and pigs show that these molecules can regulate multiple genes simultaneously, thereby offering an effective way to modify economically important yield characteristics. However, the relationships between these miRNAs, their target genes, and yield traits should be elucidated more clearly, in order to be able to apply them in breeding programs.

28.6.2 Translation of “omics” data to agriculture

The main challenge facing by today’s molecular biology community is to interpret the wealth of data generated by genome sequencing projects. Previously, molecular biology studies were performed purely in the experimental laboratory, but the huge increase in the availability of data generated in this genomic era makes computational biology essential to the research process. With the emergence of new tools and databases in molecular biology, we are now able to conduct research not only at the genome level but also at the proteome, transcriptome, and metabolome levels. The challenges facing the bioinformatic field today include the intelligent and efficient storage of the large amount of data produced, and the provision of easy, reliable, and user-friendly access to this data. Therefore streamlined computer tools should be developed to allow for the extraction of meaningful biological information (Untergasser et al., 2007). Many bioinformatic resources and tools are currently available; while some are integrated packages available from commercial vendors, the large majority are developed for specific tasks by open-source projects such as Bioconductor, BioPerl, and BioPython and are available free of charge from repositories such as GitHub and CPAN.

Bioinformatics can be used to collect and store plant genetic resources, to produce stronger, disease- and insect-resistant crops, and to improve the quality of livestock, making them healthier, more resistant to disease, and more efficient. It is used in three areas of molecular biology research: molecular sequence analysis, molecular structure analysis, and molecular function analysis. Bioinformatic tools are required to study system-wide applications such as genomics (genotyping), epigenomics (histone/DNA methylation), transcriptomics (differential gene expression), proteomics (protein turnover and interactions), and metabolomics. HTP genotyping platforms used to screen variations such as SNPs, INDELS, and SSRs (see Section 28.3) can be combined with different omics technologies to generate genetic maps, QTLs, differential gene expression QTLs (eQTLs), differential methylation QTLs (mQTLs), GWAS, and diversity analysis and to develop novel molecular markers. Many of these applications are at the cutting edge of current knowledge of molecular and cell biology, and bioinformatic resources focusing on them constitute a toolbox for breeders (Shafi & Zahoor, 2019).

28.6.3 Bioinformatic resources for sustainable crop and livestock production

As noted in Section 28.2, many databases exist to collate data relevant to sustainable crop production. For example, the following are specifically focused on biotic and abiotic stress responses:

1. *Plant Stress Gene Database*: information on genes that are active in plant stress conditions (Prabha, Ghosh, & Singh, 2012).
2. *Stress-responsive Transcription Factor Database*: a comprehensive collection of *Arabidopsis thaliana* and *Oryza sativa* L. biotic and abiotic stress-responsive genes that can recognize potential locations for transcription factor-binding sites in their promoters (Shameer et al., 2009).
3. *The Global Pest and Disease Database*: online exotic pest information archive designed to support Animal and Plant Health Inspection Service programs. The database incorporates information from various sources into a single

repository pertaining to pest taxonomy, identification, biology, distribution, hosts, significance, detection, and control (<https://www.gpdd.info/>).

4. *Plant Environmental Stress Transcript Database*: for transcripts with annotated tentative orthologs from crop abiotic stress transcripts (Balaji, Crouch, Petite, & Hoisington, 2006).
5. *PhytAMP*: a database of plant natural antimicrobial peptides (Hammami, Ben Hamida, Vergoten, & Fliss, 2009).

In a similar way, livestock animal genomics has the aim of identifying the genetic and molecular origin of all animal biological processes, so that new livestock races may be developed efficiently and with minimal costs. Animals with stress-resistance traits are of particular importance, combined with good reproduction and breeding characteristics. Genomics data can help design estimation processes for animal health and also potentially be part of the breeding decision management system (Wickham, 2013). As with crops, the main objectives of animal bioinformatics are to make all sequence data available to the public through various data repositories; make clear ontological definitions of genes, proteins, and phenotypes available; and demonstrate interactions between animals and other organisms in nature. Database development and management of animal genetic resources is a necessary task to characterize, use, and protect these irreplaceable resources (Mitra & Acharya, 2003). Generally, a database of animal genetic resources is kept on a region/country basis, storing information about the breeds of various animal and poultry species in the region. It also stores data on pedigrees of various livestock species, breeds present on breeding farms, and documents physical, production, and reproductive characteristics of each breed. Socioeconomic information on farmers raising the breed is also an important component of these databases. Various databases of animal genetic resources are available; for example, the FAO hosts DAD-IS, the Domestic Animal Diversity Information System (<http://www.fao.org/animal-genetics/breed-database/dad-is/en/>). A database on genetic characterization of animal genetic resources is also being developed. This method avoids many methodological barriers in the analyses of animal genetic knowledge and thus provides a forum to turn findings for animal breeding studies into predictive models (Baurley, Perbangsa, Subagyo, & Pardamean, 2013).

As with crops a number of species-specific bioinformatic resources have been developed for farm animals, such as The Goat and Sheep EST Database (GoSh) (Caprera et al., 2007), The Pig Genome Database (PiGenome) (Lim et al., 2009), Sheep QTL/Associations Data Summary (Sheep QTLdb) (Hu, Park, & Reecy, 2016), Resource for Chicken Gene Expression (Chickspress) (McCarthy et al., 2019), and The Bovine Genome Database (BGD) (Elsik et al., 2016).

28.7 Conclusion and future prospects

With the many challenges facing agriculture in the 21st century—growing world population, resource limitation, emerging pests and pathogens, and climate change—it is clear that effective deployment of the genetic resources provided in the natural world is essential. For example, leading agricultural scientists have recently proposed a “5G” strategy to sustain future breeding programs: genome assembly, germplasm characterization, gene function identification, genomic selection, and gene editing (Varshney et al., 2020). Bioinformatics plays an essential role in all of these technologies, and while this chapter highlights many excellent resources already available, considerable work remains to be done. Three areas in particular that require urgent development are:

1. Genomic data resources for locally adapted and orphan crop and livestock species to be created, and brought up to the same standard as those for elite crops (Section 28.2).
2. HTP genotyping technologies (Section 28.4) remain prohibitively expensive for lower income countries, making it difficult for the development of local crops to compete with global varieties. Bioinformatic tools and strategies to transfer knowledge from better characterized germplasm, and optimize genotyping protocols to reduce costs, could help to redress this imbalance.
3. There remains a huge knowledge gap in identifying gene function from sequences, and in the characterizing genetic diversity in many crop or livestock populations. Studies at a population level such as GWAS (Section 28.5) collect data on phenotype as well as genotype, but only a minority can be directly linked to specific genes. Future bioinformatic resources will need to find ways to integrate indirect phenotyping data with pan-genome databases to generate functional predictions that can be tested experimentally.

In summary, although a relatively young discipline, bioinformatics has contributed greatly to agricultural science in the last two decades. Modern computing resources have allowed increasingly detailed studies of agricultural biology at the molecular level; however, many more novel tools and techniques will be needed to secure sustainable food production in the coming years.

References

- Akçaya, M. S., Bhagwat, A. A., & Cregan, P. B. (1992). Length polymorphisms of simple sequence repeat DNA in soybean. *Genetics*, *132*(4), 1131–1139.
- Akpinar, B. A., Lucas, S., & Budak, H. (2017). A large-scale chromosome-specific SNP discovery guideline. *Functional & Integrative Genomics*, *17*, 97–105.
- Akpinar, B. A., Lucas, S. J., Vrána, J., Doležel, J., & Budak, H. (2015). Sequencing chromosome 5D of *Aegilops tauschii* and comparison with its allopolyploid descendant bread wheat (*Triticum aestivum*). *Plant Biotechnology Journal*, *13*(6).
- Alaux, M., Rogers, J., Letellier, T., Flores, R., Alfama, F., Pommier, C., et al. (2018). Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. *Genome Biology*, *19*, 111.
- Allen, A. M., Barker, G. L. A., Wilkinson, P., BurrIDGE, A., Winfield, M., Coghill, J., et al. (2012). Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, *11*, 279–295.
- Allen, A. M., Winfield, M. O., BurrIDGE, A. J., Downie, R. C., Benbow, H. R., Barker, G. L. A., et al. (2017). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnology Journal*, *15*(3), 390–401.
- Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403), eaar7191.
- Arita, M., Karsch-mizrachi, I., & Cochrane, G. (2020). The international nucleotide sequence database collaboration. *Nucleic Acids Research* (3), 1–5.
- Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, *465*, 627–631.
- Aukerman, M. J., & Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *The Plant Cell*, *15*(11), 2730–2741.
- Ayalew, H., Tsang, P. W., Chu, C., Wang, J., Liu, S., Chen, C., et al. (2019). Comparison of TaqMan, KASP and rhAmp SNP genotyping platforms in hexaploid wheat. *PLoS One*, *14*(5), e0217222.
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, *3*(10), e3376.
- Balaji, J., Crouch, J. H., Petite, P. V., & Hoisington, D. A. (2006). A database of annotated tentative orthologs from crop abiotic stress transcript. *Bioinformatics*, *1*, 225–227.
- Barh, D., Khan, M. S., & Davies, E. (2015). PlantOmics: The omics of plant science. In D. Barh, M. S. Khan, & E. Davies (Eds.), *PlantOmics: The omics of plant science* (1st ed.). New Delhi: Springer.
- Bar-Hen, A., Charcosset, A., Bourgoin, M., & Guiard, J. (1995). Relationship between genetic markers and morphological traits in a maize inbred lines collection. *Euphytica*, *84*, 145–154.
- Barua, U. M., Chalmers, K. J., Hackett, C. A., Thomas, W. T. B., Powell, W., & Waugh, R. (1993). Identification of RAPD markers linked to a *Rhynchosporium secalis* resistance locus in barley using near-isogenic lines and bulked segregant analysis. *Heredity (Edinb)*, *71*(2), 177–184.
- Baumbusch, L. O., Sundal, I. K., Hughes, D. W., Galau, G. A., & Jakobsen, K. S. (2001). Efficient protocols for CAPS-based mapping in *Arabidopsis*. *Plant Molecular Biology Reporter*, *19*(2), 137–149.
- Baurley, J. W., Perbangsa, A. S., Subagyo, A., & Pardamean, B. (2013). A web application and database for agriculture genetic diversity and association studies. *International Journal of Bio-Science and Bio-Technology*, *5*(6), 33–42.
- Becker, J., Vos, P., Kuiper, M., Salamini, F., & Heun, M. (1995). Combined mapping of AFLP and RFLP markers in barley. *Molecular Genetics and Genomics*, *249*(1), 65–73.
- Beckmann, J. S., & Soller, M. (1986). Restriction fragment length polymorphisms and genetic improvement of agricultural species. *Euphytica*, *35*(1), 111–124.
- Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, *10*, 980.
- Bhattacharyya, P. N., Goswami, M. P., & Bhattacharyya, L. H. (2016). Perspective of beneficial microbes in agriculture under changing climatic scenario: A review. *Journal of Phytology*, *8*, 26–41.
- Blake, V. C., Woodhouse, M. R., Lazo, G. R., Odell, S. G., Wight, C. P., Tinker, N. A., et al. (2019). GrainGenes: Centralized small grain resources and digital platform for geneticists and breeders. *Database*, *2019*(1), baz065.
- Boopathi, N. M. (2013). *Genetic mapping and marker assisted selection: Basics, practice and benefits*.
- Boopathi, N. M., & Boopathi, N. M. (2013). Genotyping of mapping population. In N. M. Boopathi (Ed.), *Genetic mapping and marker assisted selection: Basics, practice and benefits* (1st ed., pp. 39–80). India: Springer.
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–331.
- Boukar, O., Kong, L., Singh, B. B., Murdock, L., & Ohm, H. W. (2004). AFLP and AFLP-derived SCAR markers associated with *Striga gesnerioides* resistance in cowpea. *Crop Science*, *44*(4), 1259–1264.
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics (Oxford, England)*, *23*(19), 2633–2635.
- Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. La, D'Amore, R., Allen, A. M., et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, *491*(7426), 705–710.
- Breria, C. M., Hsieh, C. H., Yen, T. B., Yen, J. Y., Noble, T. J., & Schafleitner, R. (2020). A SNP-based genome-wide association study to mine genetic loci associated to salinity tolerance in mungbean (*Vigna radiata* L.). *Genes (Basel)*, *11*(7), 759.

- Broccanello, C., Chiodi, C., Funk, A., McGrath, J. M., Panella, L., & Stevanato, P. (2018). Comparison of three PCR-based assays for SNP genotyping in plants. *Plant Methods*, 14(1), 28.
- Broccanello, C., Gerace, L., & Stevanato, P. (2020). QuantStudio™ 12K Flex OpenArray® System as a Tool for High-Throughput Genotyping and Gene Expression Analysis. *Methods in Molecular Biology*, 2065, 199–208. Available from https://doi.org/10.1007/978-1-4939-9833-3_15.
- Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12), e1002822.
- Caprera, A., Lazzari, B., Stella, A., Merelli, I., Caetano, A. R., & Mariani, P. (2007). GoSh: A web-based database for goat and sheep EST sequences. *Bioinformatics (Oxford, England)*, 23(8), 1043–1045.
- Cardon, L. R., & Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews. Genetics*, 2, 91–99.
- Cech, T. R., & Steitz, J. A. (2014). The noncoding RNA revolution - Trashing old rules to forge new ones. *Cell*, 157, 77–94.
- Cheema, J., & Dicks, J. (2009). Computational approaches and software tools for genetic linkage map estimation in plants. *Briefings in Bioinformatics*, 10(6), 595–608.
- Chellappan, P., Xia, J., Zhou, X., Gao, S., Zhang, X., Coutino, G., et al. (2010). siRNAs from miRNA sites mediate DNA methylation of target genes. *Nucleic Acids Research*, 38, 6883–6894.
- Chen, X. (2012). Small RNAs in development - Insights from plants. *Current Opinion in Genetics & Development*, 22(4), 361–367.
- Cheng, H. Y., Wang, Y., Tao, X., Fan, Y. F., Dai, Y., Yang, H., et al. (2016). Genomic profiling of exogenous abscisic acid-responsive microRNAs in tomato (*Solanum lycopersicum*). *BMC Genomics*, 17(1), 1–13.
- Chow, C. N., Lee, T. Y., Hung, Y. C., Li, G. Z., Tseng, K. C., Liu, Y. H., et al. (2019). Plantpan3.0: A new and updated resource for reconstructing transcriptional regulatory networks from chip-seq experiments in plants. *Nucleic Acids Research*, 47(D1), D1155–D1163.
- Chuang, J. C., & Jones, P. A. (2007). Epigenetics and microRNAs. *Pediatric Research*, 61, 24–29.
- Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., et al. (2009). Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, 10(1), 582.
- Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142(1–2), 169–196.
- Cordeiro, G. M., Casu, R., McIntyre, C. L., Manners, J. M., & Henry, R. J. (2001). Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Science (Shannon, Ireland)*, 160(6), 1115–1123.
- Covarrubias, A. A., & Reyes, J. L. (2010). Post-transcriptional gene regulation of salinity and drought responses by plant microRNAs. *Plant, Cell Environment*, 33, 481–489.
- Dai, X., Zhuang, Z., & Zhao, P. X. (2018). psRNATarget: A plant small RNA target analysis server (2017 release). *Nucleic Acids Research*, 46(W1), W49–W54.
- Desjardin, C., Vaiman, A., Mata, X., Legendre, R., Laubier, J., Kennedy, S. P., et al. (2014). Next-generation sequencing identifies equine cartilage and subchondral bone miRNAs and suggests their involvement in osteochondrosis physiopathology. *BMC Genomics*, 15, 798.
- de Vienne, D. (2002). In D. de Vienne (Ed.), *Molecular markers in plant genetics and biotechnology* (1st ed., p. 248). CRC Press.
- Edet, O. U., Gorafi, Y. S. A., Nasuda, S., & Tsujimoto, H. (2018). DArTseq-based analysis of genomic relationships among species of tribe Triticeae. *Scientific Reports*, 8, 16397.
- Edwards, K. J., Barker, J. H. A., Daly, A., Jones, C., & Karp, A. (1996). Microsatellite libraries enriched for several microsatellite sequences in plants. *Biotechniques*, 20(5), 758–760.
- Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., et al. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Science (Shannon, Ireland)*, 270, 123–130.
- Eldem, V., Çelikkol Akçay, U., Ozhuner, E., Bakir, Y., Uranbey, S., & Unver, T. (2012). Genome-wide identification of miRNAs responsive to drought in peach (*Prunus persica*) by high-throughput deep sequencing. *PLoS One*, 7, e50298.
- El-Halawany, N., Abdel-Shafy, H., Shawky, A. E. M. A., Abdel-Latif, M. A., Al-Tohamy, A. F. M., & Abd El-Moneim, O. M. (2017). Genome-wide association study for milk production in Egyptian buffalo. *Livestock Science*, 198, 10–16.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, 6(5), e19379.
- Elsik, C. G., Unni, D. R., Diesh, C. M., Tayal, A., Emery, M. L., Nguyen, H. N., et al. (2016). Bovine genome database: New tools for gleaning function from the *Bos taurus* genome. *Nucleic Acids Research*, 44(D1), D834–D839.
- Eujayl, I., Baum, M., Powell, W., Erskine, W., & Pehu, E. (1998). A genetic linkage map of lentil (*Lens* sp.) based on RAPD and AFLP markers using recombinant inbred lines. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 97(1–2), 83–89.
- Falker-Gieske, C., Blaj, I., Preuß, S., Bennewitz, J., Thaller, G., & Tetens, J. (2019). GWAS for meat and carcass traits using imputed sequence level genotypes in pooled F2-designs in pigs. *G3 Genes, Genomes, Genetics*, 9(9), 2823–2834.
- Fang, J., Zhu, X., Wang, C., & Shangguan, L. (2016). Applications of DNA technologies in agriculture. *Current Genomics*, 17(4), 379–386.
- FAO. (2022). <https://www.fao.org/sustainable-development-goals/overview/fao-and-the-2030-agenda-for-sustainable-development/sustainable-agriculture/en/>. (Accessed 23 February 2022).
- Ferreira, A., da Silva, M. F., da Costa e Silva, L., & Cruz, C. D. (2006). Estimating the effects of population size and type on the accuracy of genetic maps. *Genetics and Molecular Biology*, 29, 182–192.
- Flint, J., & Eskin, E. (2012). Genome-wide association studies in mice. *Nature Reviews. Genetics*, 13, 807.

- Francisco, M., Lazaruk, K. D., Rhodes, M. D., & Wenz, M. H. (2005). Assessment of two flexible and compatible SNP genotyping platforms: TaqMan[®] SNP genotyping assays and the SNPlex[™] genotyping system. *Mutation Research/Fundamental and Molecular Mechanisms Mutagen*, 573(1–2), 111–135.
- Freebern, E., Santos, D. J. A., Fang, L., Jiang, J., Parker Gaddis, K. L., Liu, G. E., et al. (2020). GWAS and fine-mapping of livability and six disease traits in Holstein cattle. *BMC Genomics*, 21(1), 41.
- Gabriel, S., Ziaugra, L., & Tabbaa, D. (2009). SNP genotyping using the Sequenom MassARRAY iPLEX Platform. *Current Protocols in Human Genetics*, 60(1), 2.12.1–2.12.18.
- Galuszynski, N. C., & Potts, A. J. (2020). *Application of high resolution melt analysis (HRM) for screening haplotype variation in non-model plants: A case study of Honeybush (Cyclopia Vent.)*. bioRxiv; 2020.02.05.921080.
- Gebreselassie, G., Liang, B., Berihulay, H., Islam, R., Abied, A., Jiang, L., et al. (2020). Genomic mapping identifies two genetic variants in the MC1R gene for coat colour variation in Chinese Tan sheep. *PLoS One*, 15(8), e0235426.
- Gim, J. A., Ayarpadikannan, S., Eo, J., Kwon, Y. J., Choi, Y., Lee, H. K., et al. (2014). Transcriptional expression changes of glucose metabolism genes after exercise in thoroughbred horses. *Gene*, 547, 152–158.
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*.
- Guan, D., Yan, B., Thieme, C., Hua, J., Zhu, H., Boheler, K. R., et al. (2017). PlaMoM: A comprehensive database compiles plant mobile macromolecules. *Nucleic Acids Research*, 45(D1), D1021–D1028.
- Guo, Z., Yang, W., Chang, Y., Ma, X., Tu, H., Xiong, F., et al. (2018). Genome-wide association studies of image traits reveal genetic architecture of drought resistance in rice. *Molecular Plant*, 11(6), 789–805.
- Gupta, B. (2014). The attributes of RNA interference in relation to plant abiotic stress tolerance. *Gene Technology*, 3(1), 1000110.
- Gupta, P. K., Roy, J. K., & Prasad, M. (2001). Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science*, 80(4), 524–535.
- Haanstra, J. P. W., Wye, C., Verbakel, H., Meijer-Dekens, F., Van den Berg, P., Odinet, P., et al. (1999). An integrated high-density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum* × *L. pennellii* F2 populations. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 99(1–2), 254–271.
- Haley, S. D., Afanador, L., & Kelly, J. D. (1994). Selection for monogenic pest resistance traits with coupling- and repulsion-phase RAPD markers. *Crop Science*, 34(4), crops1994.0011183X003400040041x.
- Hammami, R., Ben Hamida, J., Vergoten, G., & Fliss, I. (2009). PhytAMP: A database dedicated to antimicrobial plant peptides. *Nucleic Acids Research*, 37, D963–D968.
- Han, J., Fang, J., Wang, C., Yin, Y., Sun, X., Leng, X., et al. (2014). Grapevine microRNAs responsive to exogenous gibberellin. *BMC Genomics*, 15(1), 111.
- Hayward, A. C., Tollenaere, R., Dalton-Morgan, J., & Batley, J. (2015). Molecular marker applications in plants. *Methods in Molecular Biology*, 1245, 13–27.
- Hayward, J. J., Kelly-Smith, M., Boyko, A. R., Burmeister, L., de Risio, L., Mellersh, C., et al. (2020). A genome-wide association study of deafness in three canine breeds. *PLoS One*, 15(5), e0232900.
- He, C., Holme, J., & Anthony, J. (2014). SNP genotyping: The KASP assay. In D. Fleury, & R. Whitford (Eds.), *Crop breeding: Methods and protocols* (pp. 75–86). New York, NY: Springer.
- Helentjaris, T., Slocum, M., Wright, S., Schaefer, A., & Nienhuis, J. (1986). Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 72, 257–264.
- Helmstetter, A. J., Oztolan-Erol, N., Lucas, S. J., & Buggs, R. J. A. (2020). Genetic diversity and domestication of hazelnut (*Corylus avellana* L.) in Turkey. *Plants, People, Planet*, 2(4), 326–339.
- Hiremath, P. J., Kumar, A., Penmetsa, R. V., Farmer, A., Schlueter, J. A., Chamarthi, S. K., et al. (2012). Large-scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. *Plant Biotechnology Journal*, 10(6), 716–732.
- Hou, J., An, X., Song, Y., Cao, B., Yang, H., Zhang, Z., et al. (2017). Detection and comparison of microRNAs in the caprine mammary gland tissues of colostrum and common milk stages. *BMC Genetics*, 18(1), 38.
- Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., et al. (2020). Ensembl genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Research*, 48(D1), D689–D695.
- Hu, J., & Vick, B. A. (2003). Target region amplification polymorphism: A novel marker technique for plant genotyping. *Plant Molecular Biology Reporter*, 21(3), 289–294.
- Hu, Z. L., Park, C. A., & Reecy, J. M. (2016). Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Research*, 44(D1), D827–D833.
- Huang, H. Y., Lin, Y. C. D., Li, J., Huang, K. Y., Shrestha, S., Hong, H. C., et al. (2020). MiRTarBase 2020: Updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Research*.
- Huang, S. Q., Peng, J., Qiu, C. X., & Yang, Z. M. (2009). Heavy metal-regulated new microRNAs from rice. *Journal of Inorganic Biochemistry*, 103(2), 282–287.
- Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, 44(1), 32–39.
- Hussain, B., Lucas, S. J., Ozturk, L., & Budak, H. (2017). Mapping QTLs conferring salt tolerance and micronutrient concentrations at seedling stage in wheat. *Scientific Reports*, 7(1).

- Hwang, T.-Y., Sayama, T., Takahashi, M., Takada, Y., Nakamoto, Y., Funatsuki, H., et al. (2009). High-density integrated linkage map based on SSR markers in soybean. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 16(4), 213–225.
- Ilic, K., Berleth, T., & Provart, N. J. (2004). BlastDigester – A web-based program for efficient CAPS marker design. *Trends in Genetics: TIG*, 20(7), 280–283.
- Integrated DNA Technologies I. rhAmp SNP Genotyping System [Internet]. (2020). <<https://eu.idtdna.com/pages/products/qpcr-and-pcr/genotyping/rhamp-snp-genotyping>>.
- Jaccoud, D., Peng, K., Feinstein, D., & Kilian, A. (2001). Diversity Arrays: A solid state technology for sequence information independent genotyping. *Nucleic Acids Research*, 29(4), e25.
- Jena, K. K., & Mackill, D. J. (2008). Molecular markers and their use in marker-assisted selection in rice. *Crop Science*, 48(4), 1266–1276.
- Jevsinek Skok, D., Godnic, I., Zorc, M., Horvat, S., Dovc, P., Kovac, M., et al. (2013). Genome-wide in silico screening for microRNA genetic variability in livestock species. *Animal Genetics*, 44(6), 669–677.
- Joobeur, T., Periam, N., de Vicente, M. C., King, G. J., & Arús, P. (2000). Development of a second generation linkage map for almond using RAPD and SSR markers. *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, 43(4), 649–655.
- Julio, E., Verrier, J. L., De., & Borne, F. D. (2006). Development of SCAR markers linked to three disease resistances based on AFLP within *Nicotiana tabacum* L. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 112(2), 335–346.
- Jung, S., Lee, T., Cheng, C. H., Buble, K., Zheng, P., Yu, J., et al. (2019). 15 years of GDR: New data and functionality in the genome database for Rosaceae. *Nucleic Acids Research*, 47(Database Issue), D1137–D1145.
- Kantar, M., Lucas, S. J., & Budak, H. (2011). miRNA expression patterns of *Triticum dicoccoides* in response to shock drought stress. *Planta*, 233, 471–484.
- Kim, J., Manivannan, A., Kim, D.-S., Lee, E.-S., & Lee, H.-E. (2019). Transcriptome sequencing assisted discovery and computational analysis of novel SNPs associated with flowering in *Raphanus sativus* in-bred lines for marker-assisted backcross breeding. *Horticulture Research*, 6(1), 120.
- Kofler, R., Schlötterer, C., & Lelley, T. (2007). SciRoKo: A new tool for whole genome microsatellite search and investigation. *Bioinformatics*, 1683–1685.
- Konieczny, A., & Ausubel, F. M. (1993). A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *The Plant Journal: For Cell and Molecular Biology*, 4(2), 403–410.
- Kooke, R., Wijnker, E., & Keurentjes, J. J. B. (2012). Backcross populations and near isogenic lines. *Methods in Molecular Biology*, 871, 3–16.
- Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, 9, 29. Available from <https://doi.org/10.1186/1746-4811-9-29>.
- Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: From microRNA sequences to function. *Nucleic Acids Research*, 47(D1), D155–D162.
- Kremer, A., Abbott, A. G., Carlson, J. E., Manos, P. S., Plomion, C., Sisco, P., et al. (2012). Genomics of *Fagaceae*. *Tree Genetics & Genomes*, 8, 583–610.
- Ku, H. Y., & Lin, H. (2014). PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. *National Science Review*, 1(2), 205–218.
- Kujur, A., Bajaj, D., Upadhyaya, H. D., Das, S., Ranjan, R., Shree, T., et al. (2015). Employing genome-wide SNP discovery and genotyping strategy to extrapolate the natural allelic diversity and domestication patterns in chickpea. *Frontiers in Plant Science*, 6, 162.
- Kumar, V., Singh, A., Mithra, S. V. A., Krishnamurthy, S. L., Parida, S. K., Jain, S., et al. (2015). Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 22(2), 133–145.
- Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75, 843–854.
- Levy-Sakin, M., & Ebenstein, Y. (2013). Beyond sequencing: Optical mapping of DNA in the age of nanotechnology and nanoscopy. *Current Opinion in Biotechnology*, 24(4), 690–698.
- LGC_Biosearch_Technologies. (2020). *The Array Tape[®] platform: Paradigm changing solutions*. <<https://www.douglasscientific.com/Technology/Default.aspx>>.
- Li, B., Chen, L., Sun, W., Wu, D., Wang, M., Yu, Y., et al. (2020). Phenomics-based GWAS analysis reveals the genetic architecture for drought resistance in cotton. *Plant Biotechnology Journal*, 18(12), 2533–2544.
- Li, B., Qin, Y., Duan, H., Yin, W., & Xia, X. (2011). Genome-wide characterization of new and drought stress responsive microRNAs in *Populus euphratica*. *Journal of Experimental Botany*, 62(11), 3765–3779.
- Li, G., & Quiros, C. F. (2001). Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: Its application to mapping and gene tagging in *Brassica*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 103(2–3), 455–461.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079.
- Li, J., Dai, X., Zhuang, Z., & Zhao, P. X. (2016). LegumeIP 2.0 - A platform for the study of gene function and genome evolution in legumes. *Nucleic Acids Research*, 44(D1), D1189–D1194.
- Lim, D., Cho, Y. M., Lee, K. T., Kang, Y., Sung, S., Nam, J., et al. (2009). The pig genome database (PiGenome): An integrated database for pig genome research. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 20(1), 60–66.
- Liu, D., Song, Y., Chen, Z., & Yu, D. (2009). Ectopic expression of miR396 suppresses GRF target gene expression and alters leaf growth in *Arabidopsis*. *Physiologia Plantarum*, 136, 223–236.

- Lu, X. M., Hu, X. J., Zhao, Y. Z., Song, W. B., Zhang, M., Chen, Z. L., et al. (2012). Map-based cloning of *zb7* encoding an IPP and DMAPP synthase in the MEP pathway of maize. *Molecular Plant*, 5(5), 1100–1112.
- Luan, M., Xu, M., Lu, Y., Zhang, L., Fan, Y., & Wang, L. (2015). Expression of *zma-miR169* miRNAs and their target *ZmNF-YA* genes in response to abiotic stress in maize leaves. *Gene*, 555(2), 178–185.
- Lucas, S. J., Akpinar, B. A., Kantar, M., Weinstein, Z., Aydinoglu, F., Šafář, J., et al. (2013). Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PLoS One*, 8.
- Lucas, S. J., & Budak, H. (2012). Sorting the wheat from the Chaff: Identifying miRNAs in genomic survey sequences of *Triticum aestivum* chromosome 1AL. *PLoS One*, 7(7), e40859.
- Lucas, S. J., Salantur, A., Yazar, S., & Budak, H. (2017). High-throughput SNP genotyping of modern and wild emmer wheat for yield and root morphology using a combined association and linkage analysis. *Functional & Integrative Genomics*, 17(6).
- Lucas, S. J., Šimková, H., Šafář, J., Jurman, I., Cattonaro, F., Vautrin, S., et al. (2012). Functional features of a single chromosome arm in wheat (1AL) determined from its structure. *Functional & Integrative Genomics*, 12(1), 173–182.
- Mallory, A. C., Bartel, D. P., & Bartel, B. (2005). MicroRNA-directed regulation of *Arabidopsis* auxin response Factor17 is essential for proper development and modulates expression of early auxin response genes. *The Plant Cell*, 17(5), 1360–1375.
- Mangini, G., Gadaleta, A., Colasuonno, P., Marcotuli, I., Signorile, A. M., Simeone, R., et al. (2018). Genetic dissection of the relationships between grain yield components by genome-wide association mapping in a collection of tetraploid wheats. *PLoS One*, 13, e0190162.
- Mason, A. S. (2015). *SSR genotyping. Plant genotyping* (pp. 77–89). Springer.
- Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., et al. (2004). Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nature Methods*, 1(2), 109–111.
- Mayer, K. F. X., Rogers, J., Dole el, J., Pozniak, C., Eversole, K., Feuillet, C., et al. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, 345(6194), 1251788–1251788.
- McCarthy, F. M., Pendarvis, K., Cooksey, A. M., Gresham, C. R., Bomhoff, M., Davey, S., et al. (2019). Chickspress: A resource for chicken gene expression. Database. 2019 Jan 1;baz058.
- Medina, C. A., Hawkins, C., Liu, X. P., Peel, M., & Yu, L. X. (2020). Genome-wide association and prediction of traits related to salt tolerance in autotetraploid alfalfa (*Medicago sativa* L.). *International Journal of Molecular Sciences*, 21(9), 3361.
- Meksem, K., & Kahl, G. (2005). *The handbook of plant genome mapping: Genetic and physical mapping*.
- Mendisco, F., Keyser, C., Hollard, C., Seldes, V., Nielsen, A. E., Crubézy, E., et al. (2011). Application of the iPLEXTM Gold SNP genotyping method for the analysis of Amerindian ancient DNA samples: Benefits for ancient population studies. *Electrophoresis*, 32(3-4), 386–393.
- Michelmore, R. W., Paran, I., & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *International Journal of Molecular Sciences*, 88(21), 9828–9832.
- Mitra, S., & Acharya, T. (2003). ©2003. Hoboken, NJ: John Wiley & Sons, Ltd (10.1111) *Data mining: Multimedia, soft computing and bioinformatics* (1st ed., p. 424) New York, NY: John Wiley & Sons, Inc.
- Mobuchon, L., Marthey, S., Boussaha, M., Le Guillou, S., Leroux, C., & Le Provost, F. (2015). Annotation of the goat genome using next generation sequencing of microRNA expressed by the lactating mammary gland: Comparison of three approaches. *BMC Genomics*, 16, 285.
- Morris, K. V., & Mattick, J. S. (2014). The rise of regulatory RNA. *Nature Reviews. Genetics*, 15, 423–437.
- Murray, J. D., & Anderson, G. B. (2000). Genetic engineering and cloning may improve milk, livestock production. *California Agriculture*, 54(4), 57–65.
- Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., et al. (2018). DNA molecular markers in plant breeding: Current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment*, 32(2), 261–285.
- Naqvi, N. I., & Chattoo, B. B. (1996). Development of a sequence characterized amplified region (SCAR) based indirect selection method for a dominant blast-resistance gene in rice. *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, 39(1), 26–30.
- Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., Estelle, M., et al. (2006). A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science*, 312(5772), 436–439.
- Neff, M. M., Neff, J. D., Chory, J., & Pepper, A. E. (1998). dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: Experimental applications in *Arabidopsis thaliana* genetics. *The Plant Journal: For Cell and Molecular Biology*, 14(3), 387–392.
- Neff, M. M., Turk, E., & Kalishman, M. (2002). Web-based primer design for single nucleotide polymorphism analysis. *Trends in Genetics: TIG*, 18(12), 613–615.
- Olsen, H. G., Hayes, B. J., Kent, M. P., Nome, T., Svendsen, M., Larsgard, A. G., et al. (2011). Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12. *Animal Genetics*, 42(5), 466–474.
- Olson, M., Hood, L., Cantor, C., & Botstein, D. (1989). A common language for physical mapping of the human genome. *Science*, 245(4925), 1434–1435.
- O'Rourke, J. A. (2014). Genetic and physical map correlation. *eLS*.
- Osuna-Cruz, C. M., Paytuyvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Cigliano, R. A., et al. (2018). PRGdb 3.0: A comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Research*, 46(D1), D1197–D1201.
- Ozturk, S. C., Ozturk, S. E., Celik, I., Stampar, F., Veberic, R., Doganlar, S., et al. (2017). Molecular genetic diversity and association mapping of nut and kernel traits in Slovenian hazelnut (*Corylus avellana*) germplasm. *Tree Genetics and Genomes*, 13.

- Paran, I., & Michelmore, R. W. (1993). Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 85(8), 985–993.
- Paterson, A. H. (2009). Plant genome mapping: Strategies and applications. In H. W. Doelle, S. Rokem, & M. Berovic (Eds.), *Biotechnology: Fundamentals in biotechnology* (pp. 291–315). Oxford: Eolss Publishers.
- Patisthan, J., Hartley, T. N., Fonseca de Carvalho, R., & Maathuis, F. J. M. (2018). Genome-wide association studies to identify rice salt-tolerance markers. *Plant, Cell & Environment*, 41(5), 970–982.
- Paulsmeyer, M. N., Brown, P. J., & Juvik, J. A. (2018). Discovery of anthocyanin Acyltransferase1 (AAT1) in maize using genotyping-by-sequencing (GBS). *G3 Genes|Genomes|Genetics*, 8(11), 3669.
- Pellicer, J., & Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. *The New Phytologist*, 226(2), 301–305.
- Peng, J., Zhao, J. S., Shen, Y. F., Mao, H. G., & Xu, N. Y. (2015). MicroRNA expression profiling of lactating mammary gland in divergent phenotype swine breeds. *International Journal of Molecular Sciences*, 16(1), 1448–1465.
- Perkel, J. (2008). SNP genotyping: Six technologies that keyed a revolution. *Nature Methods*, 5(5), 447–453.
- Peterson, G. W., Dong, Y., Horbach, C., & Fu, Y.-B. (2014). Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity*, 6(4), 665–680.
- Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome*, 5(3).
- Prabha, R., Ghosh, I., & Singh, D. P. (2012). Plant stress gene database: A collection of plant genes responding to stress condition. *ARPN Journal of Science and Technology*, 1(1), 28–31.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.
- Rajesh, M. K., Jerard, B. A., Preethi, P., Thomas, R. J., Fayas, T. P., Rachana, K. E., et al. (2013). Development of a RAPD-derived SCAR marker associated with tall-type palm trait in coconut. *Scientia Horticulturae*, 150, 312–316.
- Ranc, N., Muñoz, S., Xu, J., Le Paslier, M. C., Chauveau, A., Bounon, R., et al. (2012). Genome-Wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var. cerasiforme. *G3 Genes, Genomes, Genetics*, 2(8), 853–864.
- Remita, M. A., Lord, E., Agharbaoui, Z., Leclercq, M., Badawi, M., Makarenkov, V., et al. (2016). WMP: A novel comprehensive wheat miRNA database, including related bioinformatics software. *Current Plant Biology*, 7(8), 31–33.
- Rimbert, H., Darrier, B., Navarro, J., Kitt, J., Choulet, F., Leveugle, M., et al. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS One*, 13(1), e0186329.
- Robinson, A. J., Love, C. G., Batley, J., Barker, G., & Edwards, D. (2004). Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics (Oxford, England)*, 20(9), 1475–1476.
- Rogers, K., & Chen, X. (2013). Biogenesis, turnover, and mode of action of plant microRNAs. *The Plant Cell*, 25, 2383–2399.
- Roose, M. L. (2007). Mapping and marker-assisted selection. In I. A. Khan (Ed.), *Citrus genetics, breeding and biotechnology* (1st ed., pp. 275–286). CAB International.
- Rostoks, N., Ramsay, L., MacKenzie, K., Cardle, L., Bhat, P. R., Roose, M. L., et al. (2006). Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 18656–18661.
- Russell, J. R., Fuller, J. D., Macaulay, M., Hatz, B. G., Jahoor, A., Powell, W., et al. (1997). Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 95(4), 714–722.
- Scheben, A., Batley, J., & Edwards, D. (2018). Revolution in genotyping platforms for crop improvement. *Advances in Biochemical Engineering/Biotechnology*.
- Schmidt, R. (2000). Synteny: Recent advances and future prospects. *Current Opinion in Plant Biology*, 3(2), 97–102.
- Schneider, K. (2005). Mapping populations and principles of genetic mapping. *The Handbook of Plant Genome Mapping: Genetic and Physical Mapping*, 3–19.
- Scott, M. F., Ladejobi, O., Amer, S., Bentley, A. R., Biernaskie, J., Boden, S. A., et al. (2020). Multi-parent populations in crops: A toolbox integrating genomics and genetic mapping with breeding. *Heredity (Edinb)*, 125, 396–416.
- Semagn, K., Babu, R., Hearne, S., & Olsen, M. (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement. *Molecular Breeding*, 33(1), 1–14.
- Semagn, K., Bjørnstad, A., & Ndjioudjop, M. N. (2006). Principles, requirements and prospects of genetic mapping in plants. *African Journal of Biotechnology*, 5(25), 2569–2587.
- Shafi, A., & Zahoor, I. (2019). Bioinformatics and plant stress management. In K. R. Hakeem, N. A. Shaik, B. Banaganapalli, & R. Elango (Eds.), *Essentials of bioinformatics, volume III: In silico life sciences: Agriculture* (1st ed., pp. 47–78). Cham: Springer International Publishing.
- Shafi, U., Mumtaz, R., García-Nieto, J., Hassan, S. A., Zaidi, S. A. R., Iqbal, N. (2019). Precision agriculture techniques and practices: From considerations to applications. *Sensors (Switzerland)*.
- Shakoor, N., Lee, S., & Mockler, T. C. (2017). High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Current Opinion in Plant Biology*, 38, 184–192.
- Shameer, K., Ambika, S., Varghese, S. M., Karaba, N., Udayakumar, M., & Sowdhamini, R. (2009). STIFDB arabidopsis stress responsive transcription factor database. *International Journal of Plant Genomics*, 2009, 583429.
- Shao, N., Jiang, S. M., Zhang, M., Wang, J., Guo, S. J., Li, Y., et al. (2014). MACRO: A combined microchip-PCR and microarray system for high-throughput monitoring of genetically modified organisms. *Analytical Chemistry*, 86(2), 1269–1276.

- Shelef, O., Weisberg, P. J., & Provenza, F. D. (2017). The value of native plants and local production in an era of global agriculture. *Frontiers in Plant Science*, 8, 2069.
- Shen, R., Fan, J.-B., Campbell, D., Chang, W., Chen, J., Doucet, D., et al. (2005). High-throughput SNP genotyping on universal bead arrays. *Mutation Research - Fundamental and Molecular Mechanisms Mutagen*, 573(1), 70–82.
- Shi, A., Chen, P., Li, D., Zheng, C., Zhang, B., & Hou, A. (2009). Pyramiding multiple genes for resistance to soybean mosaic virus in soybean using molecular markers. *Molecular Breeding*, 23(1), 113.
- Siddique, M. I., Lee, H.-Y., Ro, N.-Y., Han, K., Venkatesh, J., Solomon, A. M., et al. (2019). Identifying candidate genes for *Phytophthora capsici* resistance in pepper (*Capsicum annuum*) via genotyping-by-sequencing-based QTL mapping and genome-wide association study. *Scientific Reports*, 9(1), 9962.
- Simko, I. (2016). High-resolution DNA melting analysis in plant research. *Trends in Plant Science*, 21(6), 528–537.
- Słomka, M., Sobalska-Kwapis, M., Wachulec, M., Bartosz, G., & Strapagiel, D. (2017). High resolution melting (HRM) for high-throughput genotyping—Limitations and caveats in practical case studies. *International Journal of Molecular Sciences*, 18(11), 2316.
- Spannagl, M., Nussbaumer, T., Bader, K. C., Martis, M. M., Seidel, M., Kugler, K. G., et al. (2016). PGSB plantsDB: Updates to the database framework for comparative plant genome research. *Nucleic Acids Research*, 44(D1), D1141–D1147.
- Stemmers, F. J., & Gunderson, K. L. (2007). Whole genome genotyping technologies on the BeadArray™ platform. *Biotechnology Journal*, 2(1), 41–49.
- Sun, F., Liu, J., Hua, W., Sun, X., Wang, X., & Wang, H. (2016). Identification of stable QTLs for seed oil content by combined linkage and association mapping in *Brassica napus*. *Plant Science (Shannon, Ireland)*, 252, 388–399.
- Sunkar, R., Chinnusamy, V., Zhu, J., & Zhu, J. K. (2007). Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. *Trends in Plant Science*, 12(7), 301–309.
- Tamura, Y., Hattori, M., Yoshioka, H., Yoshioka, M., Takahashi, A., Wu, J., et al. (2014). Map-based cloning and characterization of a brown planthopper resistance gene BPH26 from *Oryza sativa* L. ssp. indica cultivar ADR52. *Scientific Reports*, 4, 5872.
- Tanksley, S. D., Young, N. D., Paterson, A. H., & Bonierbale, M. W. (1989). RFLP mapping in plant breeding: New tools for an old science. *Bio/Technology*, 7(3), 257–264.
- Tardieu, F. (2012). Any trait or trait-related allele can confer drought tolerance: Just design the right drought scenario. *Journal of Experimental Botany*, 63(1), 25–31.
- Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, 17(16), 6463–6471.
- Tello-Ruiz, M. K., Naithani, S., Stein, J. C., Gupta, P., Campbell, M., Olson, A., et al. (2018). Gramene 2018: Unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Research*, 46(D1), D1181–D1189.
- Thiel, T., Kota, R., Grosse, I., Stein, N., & Graner, A. (2004). SNP2CAPS: A SNP and INDEL analysis tool for CAPS marker development. *Nucleic Acids Research*, 32(1), e5.
- Thompson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding and Biotechnology*, 2(3), 195–212.
- Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., & Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics*, 28(3), 286–289.
- Tomari, Y., Du, T., Haley, B., Schwarz, D. S., Bennett, R., Cook, H. A., et al. (2004). RISC assembly defects in the *Drosophila* RNAi mutant armitage. *Cell*, 116, 831–841.
- Tsai, H. Y., Janss, L. L., Andersen, J. R., Orabi, J., Jensen, J. D., Jahoor, A., et al. (2020). Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. *Scientific Reports*, 10, 3347.
- Turner, M. K., Kolmer, J. A., Pumphrey, M. O., Bulli, P., Chao, S., & Anderson, J. A. (2017). Association mapping of leaf rust resistance loci in a spring wheat core collection. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 130, 345–361.
- Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. M. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*, 35(W), W71–W74.
- Usadel, B., Schwacke, R., Nagel, A., & Kersten, B. (2012). GabiPD - The GABI primary database integrates plant proteomic data with gene-centric information. *Frontiers in Plant Science*, 3, 154.
- Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van De Peer, Y., et al. (2018). PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research*, 46(D1), D1190–D1196.
- van Doorn, R., Szemes, M., Bonants, P., Kowalchuk, G. A., Salles, J. F., Ortenberg, E., et al. (2007). Quantitative multiplex detection of plant pathogens using a novel ligation probe-based system coupled with universal, high-throughput real-time PCR on OpenArrays™. *BMC Genomics*, 8(1), 276.
- Varshney, R. K., Marcel, T. C., Ramsay, L., Russell, J., Röder, M. S., Stein, N., et al. (2007). A high density barley microsatellite consensus map with 775 SSR loci. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 114(6), 1091–1103.
- Varshney, R. K., Mir, R. R., Bhatia, S., Thudi, M., Hu, Y., Azam, S., et al. (2014). Integrated physical, genetic and genome map of chickpea (*Cicer arietinum* L.). *Functional & Integrative Genomics*, 14, 59–73.
- Varshney, R. K., Sinha, P., Singh, V. K., Kumar, A., Zhang, Q., & Bennetzen, J. L. (2020). 5Gs for crop genetic improvement. *Current Opinion in Plant Biology*, 56, 190–196.
- Virk, P. S., Ford-Lloyd, B. V., Jackson, M. T., & Newbury, H. J. (1995). Use of RAPD for the study of diversity within plant germplasm collections. *Heredity (Edinb)*, 74(2), 170–179.

- Volante, A., Tondelli, A., Desiderio, F., Abbruscato, P., Menin, B., Biselli, C., et al. (2020). Genome wide association studies for japonica rice resistance to blast in field and controlled conditions. *Rice*, *13*, 71.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Van De Lee, T., Hornes, M., et al. (1995). AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research*, *23*(21), 4407–4414.
- Vos, P., & Zabeau, M. (1992). Selective restriction fragment amplification: A general method for DNA fingerprinting. Office EP, editor. Vol. EP0534858B. Keygene NV.
- Vuylsteke, M., Mank, R., Antonise, R., Bastiaans, E., Senior, M. L., Stuber, C. W., et al. (1999). Two high-density AFLP[®] linkage maps of *Zea mays* L.: Analysis of distribution of AFLP markers. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *99*(6), 921–935.
- Vuylsteke, M., Peleman, J. D., & Van Eijk, M. J. (2007). AFLP technology for DNA fingerprinting. *Nature Protocols*, *2*(6), 1387.
- Wang, J., Chuang, K., Ahluwalia, M., Patel, S., Umblas, N., Mirel, D., et al. (2005). High-throughput SNP genotyping by single-tube PCR with Tm-shift primers. *Biotechniques*, *39*(6), 885–893.
- Wang, J., Lin, M., Crenshaw, A., Hutchinson, A., Hicks, B., Yeager, M., et al. (2009). High-throughput single nucleotide polymorphism genotyping using nanofluidic dynamic arrays. *BMC Genomics*, *10*(1), 561.
- Wang, W., Hu, Y., Sun, D., Staehelin, C., Xin, D., & Xie, J. (2012). Identification and evaluation of two diagnostic markers linked to *Fusarium wilt* resistance (race 4) in banana (*Musa* spp.). *Molecular Biology Reports*, *39*(1), 451–459.
- Warmerdam, S., Sterken, M. G., van Schaik, C., Oortwijn, M. E. P., Sukarta, O. C. A., Lozano-Torres, J. L., et al. (2018). Genome-wide association mapping of the architecture of susceptibility to the root-knot nematode *Meloidogyne incognita* in *Arabidopsis thaliana*. *The New Phytologist*, *218*(2), 724–737.
- Waugh, R., & Powell, W. (1992). Using RAPD markers for crop improvement. *Trends in Biotechnology*, *10*, 186–191.
- Welsh, J., & McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, *18*(24), 7213–7218.
- Wenguang, Z., Jianghong, W., Jinquan, L., & Yashizawa, M. (2007). A subset of skin-expressed microRNAs with possible roles in goat and sheep hair growth based on expression profiling of mammalian microRNAs. *OMICS: A Journal of Integrative Biology*, *11*(4), 385–396.
- White, P. S., & Matisse, T. C. (2001). Genomic mapping and mapping databases. (10.1111) In (2nd ed. A. X. Baxevanis, & B. F. F. Ouellette (Eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (Vol. 43 New York, NY: John Wiley & Sons, Ltd.
- Wickham, B. (2013). Information system technology for integrated animal identification, traceability and performance recording: The example of the Irish cattle sector. *ICAR, Technical Services*, *15*, 183–195.
- Wijnhoven, B. P. L., Michael, M. Z., & Watson, D. I. (2007). MicroRNAs and cancer. *The British Journal of Surgery*, *94*, 23–30.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). Comment: The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.
- Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, *18*(22), 6531–6535.
- Winfield, M. O., Allen, A. M., Burrige, A. J., Barker, G. L. A., Benbow, H. R., Wilkinson, P. A., et al. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnology Journal*, *14*(5), 1195–1206.
- Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., et al. (2010). DNA methylation mediated by a MicroRNA Pathway. *Molecular Cell*, *38*, 465–475.
- Xiao, N., Gao, Y., Qian, H., Gao, Q., Wu, Y., Zhang, D., et al. (2018). Identification of genes related to cold tolerance and a functional allele that confers cold tolerance. *Plant Physiology*, *177*(3), 1108–1123.
- Xu, H., Wang, X., Du, Z., & Li, N. (2006). Identification of microRNAs from different tissues of chicken embryo and adult chicken. *FEBS Letters*, *580*, 3610–3616.
- Xue, J., Zhao, S., Liang, Y., Hou, C., & Wang, J. (2008). *Bioinformatics and its applications in agriculture*. IFIP International Federation for Information Processing (pp. 977–982). Boston, MA: Springer.
- Yang, C., Li, D., Mao, D., Liu, X., Ji, C., Li, X., et al. (2013). Overexpression of microRNA319 impacts leaf morphogenesis and leads to enhanced cold tolerance in rice (*Oryza sativa* L.). *Plant, Cell Environment*, *36*(12), 2207–2218.
- Yang, J., Farmer, L. M., Agyekum, A. A. A., & Hirschi, K. D. (2015). Detection of dietary plant-based small RNAs in animals. *Cell Research*.
- Yu, H.-J., Jeong, Y.-M., Lee, Y.-J., Yim, B., Cho, A., & Mun, J.-H. (2020). Marker integration and development of Fluidigm/KASP assays for high-throughput genotyping of radish. *Horticulture, Environment, and Biotechnology*, *61*(4), 767–777.
- Yu, J., & Main, D. (2015). Role of bioinformatic tools and databases in cotton research. In D. D. Fang, & R. G. Percy (Eds.), *Cotton* (Vol. 57, pp. 303–337). American Society of Agronomy.
- Yuan, C., Meng, X., Li, X., Illing, N., Ingle, R. A., Wang, J., et al. (2017). PceRBase: A database of plant competing endogenous RNA. *Nucleic Acids Research*, *45*(D), D1009–D1014.
- Zec, H. C., Zheng, T., Liu, L., Hsieh, K., Rane, T. D., Pederson, T., et al. (2018). Programmable microfluidic genotyping of plant DNA samples for marker-assisted selection. *Microsystems Nanoengineering*, *4*(1), 17097.
- Zhang, S., Yue, Y., Sheng, L., Wu, Y., Fan, G., Li, A., et al. (2013). PASmiR: A literature-curated database for miRNA molecular regulation in plant response to abiotic stress. *BMC Plant Biology*, *13*, 1–8.
- Zhang, Y., Zhang, J., Gong, H., Cui, L., Zhang, W., Ma, J., et al. (2019). Genetic correlation of fatty acid composition with growth, carcass, fat deposition and meat quality traits based on GWAS data in six pig populations. *Meat Science*, *150*, 47–55.
- Zheng, Y., Wu, S., Bai, Y., Sun, H., Jiao, C., Guo, S., et al. (2019). Cucurbit genomics database (CuGenDB): A central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Research*, *47*(D1), D1128–D1136.

- Zhigunov, A. V., Ulianich, P. S., Lebedeva, M. V., Chang, P. L., Nuzhdin, S. V., & Potokina, E. K. (2017). Development of F1 hybrid population and the high-density linkage map for European aspen (*Populus tremula* L.) using RADseq technology. *BMC Plant Biology*, *17*, 180.
- Zhou, H., Blangero, J., Dyer, T. D., Chan, Kh. K., Lange, K., & Sobel, E. M. (2017). Fast genome-wide QTL association mapping on pedigree and population data. *Genetic Epidemiology*, *41*, 174–186.
- Zhou, X., Wang, G., Sutoh, K., Zhu, J. K., & Zhang, W. (2008). Identification of cold-inducible microRNAs in plants by transcriptome analysis. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1779*(11), 780–788.
- Zietkiewicz, E., Rafalski, A., & Labuda, D. (1994). Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics*, *20*(2), 176–183.

This page intentionally left blank

Application of high-throughput structural and functional genomic technologies in crop nutrition research

Nand Lal Meena^{1,3}, Ragini Bhardwaj¹, Om Prakash Gupta², Vijay Singh Meena¹, Ajeet Singh³ and Aruna Tyagi³

¹ICAR-National Bureau of Plant Genetic Resources, New Delhi, Delhi, India, ²ICAR-Indian Institute of Wheat and Barley Research, Karnal, Haryana, India, ³Division of Biochemistry, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

29.1 Introduction

Genomics is the investigation of the genome of a life form, which is the entirety of all qualities of any person. Genomics requires the investigation of the entirety of the nucleotide successions, including primary qualities, administrative groupings, and noncoding DNA fragments, in the chromosomes of a creature. It requires assurance of the whole DNA succession of a creature, which includes around 3 billion nucleotides. The establishment for sequencing of nucleotides was laid by Fred Sanger and by Allan Maxam and Walter Gilbert. Food parts connect with our body at framework, organ, cell, and atomic levels, contingent upon their retention, bioavailability, digestion, and bioadequacy. Currently, nutritional and well-being research is focused on advancing wellbeing by delaying sickness onset and improving performance. Significantly, the helpful activity of a specific food segment at the atomic level does not cause a pernicious impact at some other level. Unraveling the atomic transaction among food and well-being requires consequently comprehensive methodologies because nourishing the improvement of certain well-being angles should not be undermined by the weakening of others. At the end of the day, in sustenance, we need to get everything right. Endless examinations are accessible which report the impact of different food parts on well-being, yet agreement as to the advantageous or impeding impacts of even a solitary segment is slippery, for different reasons, predominantly because they come up short on the framework's science approach. Applying the frameworks science approach, nutrigenomics tries to build up dietary marks that are the trademark result of an individual's supplement climate quality collaboration. In this way, sicknesses with a hereditary inclination can prompt fluctuated sorts of dietary marks, which can be analyzed at different levels, for example, cell culture, tissue culture, and entire life forms. The investigation of genomics is isolated into the accompanying two areas: structural genomics (SG) and functional genomics.

29.2 Structural genomics

SG is a global exertion to decide the three-dimensional (3D) states of extremely significant organic macromolecules, with an essential spotlight on proteins. A significant auxiliary objective is to diminish the normal expense of structure assurance through high-throughput (HT) strategies for protein creation and structure assurance. In the United States the National Institutes of Health started pilot SG ventures at nine focussing through the Protein Structure Initiative (PSI), starting in 2000. As the PSI venture moves from its pilot stage to full creation this year, the all-out subsidizing at four huge scope habitats and six particular places is required to be roughly \$60 million yearly ([The French-Italian Public Consortium for Grapevine Genome Characterization, 2007](#)). Significant assets have likewise been spent globally, with SG ventures in Japan, Canada, Israel, and Europe in progress since the last part of the 1990s. With more than 5 years of information from SG ventures around the world, this is a perfect chance to analyze their effect and to assess how much advancement has been made toward the significant objectives. Likewise, with another enormous scope, objective-based

activities, it is critical to set up target, quantitative proportions of achievement. We plan to quantify the natural significance and trouble of tackling macromolecular structures, and we depend on a few intermediaries to appraise these. Albeit each new trial structure adds to our vault of primary information, the most underlying researcher would concur with those novel structures (e.g., the main high-goal structures of ribosomal subunits are particularly important). For instance, the principal protein structure in a family might be utilized to get capacity and system, gather the overlap of other relatives, make definite near models of the most comparable proteins, or distinguish already uncharacterized developmental connections. The oddity is not restricted to new families: the structure of a formerly settled protein in alternate compliance or with an alternate restricting accomplice could give understanding into its practical instruments. The thought may likewise be given to the size, intricacy, or nature of a structure, as assessments of its trouble. Over the long haul a structure's effect on the field might be roughly assessed by the quantity of in this way distributed papers that refer to the first reference. SG is a moderately new part of the underlying science that alludes to the investigation of protein structures on a genome scale. Investigation of the atomic structure of proteins depends on the data got from genomics examination, that is, the investigation of the entirety of the qualities of a cell, tissue, or living being, at the DNA (genotype), courier RNA (mRNA; transcriptome), or protein (proteome) levels. A blend of exploratory, bioinformatics, and displaying approaches will be used to describe the structures of all proteins in a particular objective set, such as all proteins encoded by a particular genome, agents from a particular protein overlap or useful family, or even entire pathways or networks. Consequently, the improvement of HT strategies for protein structure assurance was needed to quicken progress, corresponding to that made by the genome sequencing focuses, utilizing generally X-beam crystallography and nuclear magnetic resonance (NMR) to empower thinking regarding tens to hundreds to thousands of structures as opposed to each, in turn, by customary techniques. Regularly the lone accessible data for a given objective at the beginning of an SG venture is the genomic arrangement of potential coding areas like open reading frame (ORFs), are part of a reading frame that contain no stop codon, purported "speculative proteins." Hence, SG is for the most part a disclosure-based way to deal with investigating the 3D structures of quality items, where, when all is said in done, they might be restricted to no information on the real capacity of the objective protein and regularly no dependable structure forecast can be produced using the essential succession. A general objective of SG is to reason work dependent on underlying comparability to describe proteins, fusing any data from endogenous ligands, to propose testable speculations on its particular capacity in the cell, which assists the scientist in understanding macromolecular hardware and edifices. Three helpful segments include cloning of proteins for underlying examinations, experimental techniques, computational strategies, and data investigation.

29.3 Application of structural genomics

29.3.1 To determine each single protein structure encrypted by the genome

The quantity of protein families is far more modest than the number of proteins, zeroing in the structure assurance endeavors on a couple of individuals from every family will give structure layouts suitable to the exact displaying of most relatives. The assignment is as a rule cultivated by coordinated focuses in the United States and around the world; each middle has the abilities for bioinformatics target choice, cloning of articulation vectors, HT protein cleansing, and structure assurance, either by crystallography at synchrotron sources or utilizing atomic attractive reverberation (NMR) techniques. As a feature of the PSI in the United States, a 5-year pilot stage for the undertaking has quite recently been finished with more than 1000 structures decided. The creation stage (PSI-2) began in 2005. In this stage a few communities will zero in on the creation of 200 structures for every year while others will handle more troublesome issues identifying with the assurance of the structure of eukaryotic multiarea proteins and film proteins. Since all things considered, 10,000 novel structures should be addressed to give critical inclusion of all genome successions, various long stretches of work stay for the undertaking to finish its underlying points. Be that as it may, in light of progress to date, the possibilities seem phenomenal for critical development of genome inclusion, just as gaining ground on more troublesome primary issues. Underlying genomics endeavors have generally centered on producing single protein structures of special and different targets. In any case a solitary structure for a given objective is frequently deficient to immovably dole out the capacity or to drive drug disclosure. As a component of the Seattle Structural Genomics Center for Infectious Disease, we try to grow the focal point of primary genomics by explaining troupes of structures that inspect little atom protein cooperation to choose irresistible illness targets. In this part, we talk about two applications for little atom libraries in underlying genomics: fair section screening, to give the motivation to lead improvement, and focused on, information-based screening, to affirm or address the practical explanation of a given quality item (Pires et al., 2004). This move-in accentuation brings about a primary genomics exertion that is more drawn in with the irresistible

sickness research network, and one that produces structures of more prominent utility to analysts intrigued by both protein capacity and inhibitor improvement. We likewise portray explicit strategies for directing HT piece screening in an underlying genomics setting by X-beam crystallography.

29.3.2 Identification of three-dimensional structure and folding of novel protein functions

It decides the structures of all the protein crease families encoded by the qualities of living creatures. If effective, this will permit the structures, everything being equal, or quality items to be controlled by homology to proteins where the area overlay structure has been settled. Information on a protein's structure and its homology to different proteins give experiences into the capacity of the protein and its functions inside natural frameworks. This information may permit us to regulate a protein's movement with inhibitor particles or activator atoms and by hereditary designing. All such conceivable outcomes depend on a comprehension of the protein's physical, synthetic, and mathematical properties, derived from its subatomic structure.

29.3.3 Gene and protein interactions: the role of protein structure prediction in structural genomics

It decides the structures of all the protein crease families encoded by the qualities of living creatures. On the off chance that fruitful, this will permit the structures, all things considered, or quality items to be controlled by homology to proteins where the area overlay structure has been settled. Information on a protein's structure and its homology to different proteins give bits of knowledge into the capacity of the protein and its parts inside organic frameworks. This information may permit us to adjust a protein's movement with inhibitor atoms or activator particles and by hereditary designing. All such potential outcomes depend on a comprehension of the protein's physical, compound, and mathematical properties, concluded from its atomic structure. Underlying genomics means fundamentally describing most protein arrangements by a proficient blend of trial and forecast. This point will be accomplished via a cautious choice of target proteins and their structure assurance by X-beam crystallography or NMR spectroscopy. There is an assortment of target determination plans, going from zeroing in on just novel folds to choosing all proteins in a model genome. A model-driven view necessitates that objectives be chosen to such an extent that the greater part of the excess arrangements can be demonstrated with helpful precision by near displaying. Indeed, even with underlying genomics, the structure of the vast majority of the proteins will be demonstrated, not dictated by test. As examined over, the exactness of near models and correspondingly the assortment of their applications decline strongly beneath the 30% grouping personality cutoff, fundamentally because of a fast expansion in arrangement mistakes. In this manner, we should decide protein structures so the vast majority of the excess groupings are identified within any event one known structure at higher than 30% arrangement character. It was as of late assessed that this cutoff requires at least 16,000 focuses to cover 90% of all protein area families, including those of film proteins. These 16,000 structures will permit the demonstrating of a particularly bigger number of proteins. For instance, New York Structural Genomics Research Consortium estimated the effect of its structures by recording the number and nature of the comparing models for perceptibly related proteins in the nonrepetitive grouping information base. Overall, 100 protein arrangements with no earlier primary portrayal could be demonstrated in any event at the overlay level. This huge influence of structure assurance by protein structure displaying outlines and legitimizes the reason for primary genomics. Once more structure forecast will add to primary genomics severally. Enormous scope all over again forecast can direct objective choice by zeroing in on trial structure assurance on proteins liable to embrace novel folds. Once more strategies should likewise be valuable in supplementing near displaying techniques by building parts of proteins not present in format structures. What is more, new techniques enhanced by restrictions from cross-connecting or different trials can give models to proteins not promptly agreeable to X-beam crystallographic or NMR investigation. At long last, enormous scope all over again displaying may permit coarse structure-based bits of knowledge into protein capacity of countless proteins well ahead of time of tentatively decided structures.

29.4 Dynamic expression of functional genomics

Functional genomics includes the expression of dynamic interactions of genes. It plans to relate the aggregate and genotype on genome level and incorporates cycles, for example, record, interpretation, protein–protein association, and epigenetic guideline. This includes far-reaching examination to get qualities, their practical jobs, and variable degrees of protein articulation. The practical genomics arose because of the difficulties presented by complete genome successions. To comprehend this cycle, it stands crucial toward the understanding of physio-biochemical capacity of each quality-

building elements. However, illustrations next to molecular and metabolite levels can give understanding not just into the conceivable capacity of individual quality yet besides the participation that happens among qualities and quality items to create a characterized organic result. The innovation, engaged with characterizing useful genomics, is DNA or oligonucleotide microarray innovation for deciding mRNA, 2D gels, and mass spectroscopy and different techniques for dissecting various proteins and GC–Ms or fluid chromatography–mass spectrometry (LC–Ms) for distinguishing and evaluating various metabolites in a cell. HT techniques for forward and turn around hereditary qualities are likewise basic to utilitarian genomics. Useful genomics lies on quality articulation, profiling (mRNA) in protein articulation, invert hereditary qualities, the age of focused changes in qualities of revenue other than forwarding transformation rate, the age of arbitrary changes in the genome for attractive freaks and bioinformatics. These rules help in giving the most extreme data of a specific living being. These aid in understanding the organic cycle at the atomic level and help to distinguish novel qualities directing this cycle. To comprehend the quality capacity, it is alluring to recognize qualities and to comprehend its appearance at the entire genome level. There are numerous prokaryotic and eukaryotic creatures, the genomes of which are completely sequenced. The current revelation is the planning of entire arrangements of qualities present in the human genome. It is conceivable to allot capacities to novel qualities and proteins and to comprehend organic cycles at the atomic level. The coordinated comprehension of the control of quality articulation and information on sign transduction, cell flagging, and generally cell work are dynamic instruments to contemplate the guideline of quality articulation in some random cell type. In yeast cells, records related to various stages of the cell cycle structure are of discrete groups. These examinations permitted grouping labels encoding proteins of obscure capacity to be allocated to putative classes dependent on their bunching with qualities of known capacity. Here, part of practical genomics will be to test those tedious capacities and apply them to determine complex natural cycles (Fig. 29.1).

29.5 Functional genomics approaches

1. *Transcriptomics*: Transcriptomics considers measure quality articulation at the record or RNA level, including both mRNA and ncRNAs quality articulation in a cell. Transcriptomic examinations cover the progression of passing data from DNA to RNA. As opposed to DNA, there is certainly not a solitary transcriptome yet one for every cell. What is more, it might change in various conditions.
2. *Proteomics*: Proteomics approaches center around which proteins are communicated in a natural framework yet may likewise incorporate investigations of protein structure. Proteomics research centers around protein recognizable proof, evaluation, movement, security, restriction, and capacity, which assume basic parts in cell flagging occasions (Wilkins et al., 1996). In the most recent decade, incredible advances have been accomplished in rice proteomics, which gives extensive previews on the comprehension of rice improvement, stress resilience, organelle, and protein

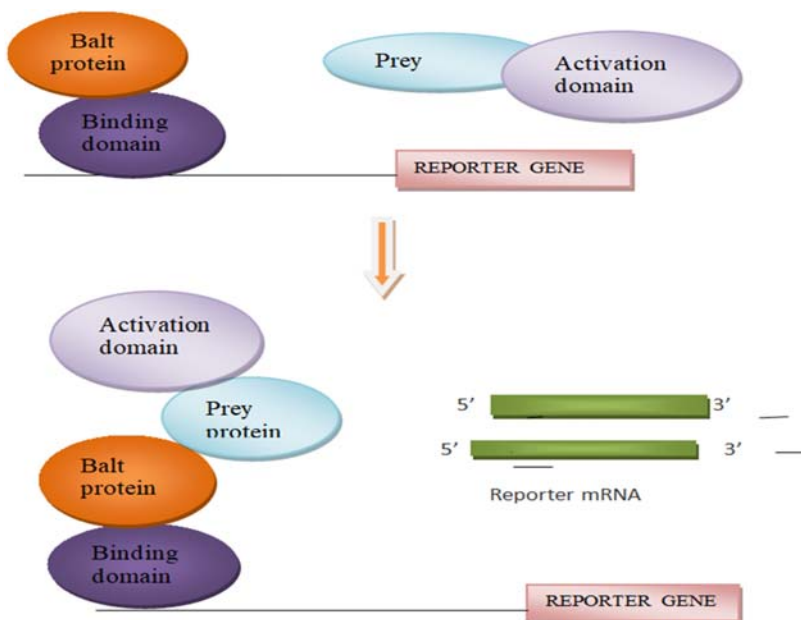


FIGURE 29.1 Functional genomics.

post-translational modification (PTM). Proteomics concentrates in rice have been performed generally utilizing gel-based (1DE, 2DE, and 2DIGE) and sans gel (LC–Ms/MS or MudPIT) approaches, and all the more as of late isobaric tags for relative and absolute quantitation (iTRAQ), for protein quantitation dependent on Ms/MS. Kim, Kim, Cho, Kim, and Kim, (2014) explored and summed up the advancement in rice proteomics concentrates from 2010 to 2013, with a significant spotlight on rice under assorted abiotic and biotic pressure conditions. All the more as of late and iTRAQ-marking-based quantitative proteomics, system was utilized to explore the proteomes under high temperature in various rice cultivars. The outcomes indicated that high-temperature stress initiated little warmth stun proteins, expansions, and lipid moves proteins in high-temperature-safe cultivars. Polyethylene glycol–mimicked dry season responsiveness in a period subordinate way in root showed that a large portion of the differentially communicated proteins gave off an impression of being associated with bioenergy and digestion (Agrawal, Devanur, & Li, 2016). By utilizing polypeptides advanced and phosphorylated by IMAC, 201 phosphopeptides indicating pistil-explicit articulation were distinguished. Protein phosphorylation is one of the most well-known posttranslational changes. It was estimated that the guideline of protein phosphorylation assumes a significant part in the development and advancement of plants. There are more than 1400 qualities that encode protein kinases and 300 qualities that encode phosphatases in rice. A mix of PolyMAC and TiO₂ advances have effectively recognized 2000 phosphoproteins from developing shame and early-stage tissues, which will enormously encourage the investigations of the turn of events and fertilization of rice disgrace (Wang, Ying, Hu, & Zhang, 2014). To additionally create proteomics and coordinate the accessible information, a few data sets of proteomics have been established a rice metaindicator of explicit phosphorylation locales (<http://bioinformatics.fafu.edu.cn/PhosphoRice/>; (Que et al., 2012)), Oryza PG-DB, a rice proteome information based on shotgun proteogenomics (<http://oryzapg.iab.keio.ac.jp/>; (Helmy, Tomita, & Ishihama, 2011)), and PRIN and anticipated rice interactome network (<http://bis.zju.edu.cn/prin/>; (Gu, Zhu, Jiao, Meng, & Chen, 2011)). Future proteomics examination should additionally refine normal and solid strategies for test readiness, including tissue reaping and protein extraction, to deliberately research the subcellular areas and posttranslational alterations of proteins and explain their natural capacities. With the advancement of neutralizer immune-protein innovation, the manufacture of protein microarrays can assist with understanding the HT recognizable proof of the useful proteins specifically natural cycles.

3. *Metabolomics*: It is the investigation of all metabolites in a natural framework and keeping in mind that these are regularly not coded for in the genome, they are delivered during cell, tissue, or living being digestion. The utilization of metabolomics in the fields of pharmacology and toxicology has prompted some achievement. Notwithstanding, most explorations in these fields have been directed on research facility creatures that are hereditarily and healthfully more homogenous than people. All the more as of late, endeavors are being had to consider the effect of supplements on the metabolome; however, such investigations must be performed on people, which makes it harder to detail a test plan that yields important information. In this manner the utilization of metabolomics to sustenance research is more confounded. Since metabolomics is the science that breaks down metabolites that are the final results that rely upon the genomics, transcriptomics, and proteomics of an individual, the metabolome speaks to the result of the supplement quality climate cooperation. Nonetheless, the examination of the little atoms that include a metabolome is no simple undertaking. Plants can create the same number of as 0.2–1 million metabolites. As of late, with the improvement of metabolomics scientific innovations, especially the development in metabolic profiling dependent on mass spectra and attractive reverberation imaging, the examination fields of metabolomics have been ceaselessly extended (Saito et al., 2013). Progress has been made in the use of plant metabolomics to the metabolite identification (MetID) of useful qualities, dismemberment of metabolic pathways, and hereditary investigation of normal varieties through the mix with other omics advancements. Conventional LC–Ms incorporates focused on and untargeted metabolomics. A foundation of metabolomics dependent on a wide range of untargeted metabolomic investigation has been set up, which can measure more than 800 known and obscure metabolites within 30 minutes (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014). The metabolomic examination of tests from 210 Recombinant inbred lines (RILs) got from a cross between two first-class indicia rice assortments, Zhenshan 97 and Minghui 63, and identified roughly 1000 metabolites, which were set out to more than 2800 metabolic quantitative trait locus (QTL) (Gong et al., 2013). Genome-wide affiliation study was utilized to distinguish a few many loci/destinations that control normal varieties in metabolite substances (Chen et al., 2014), and they clarified more than 160 new metabolites, including flavonoids, nutrients, and terpenes.
4. *Interactomics*: Interactomics is of specific pertinence to farming frameworks, especially in getting the illness. Interactomics is the investigation of the atomic communications between and includes have microbe collaborations.
5. *Nutrigenomics*: Nutrigenomics adds to exactness nourishment by disentangling the instruments of individual-to-individual and populace contrasts in light of food introductions. Multiomic inconstancies in plants and designing of

food creation are the “input” capacities prompting fluctuation in healthful results. Nutrigenomics (or “healthful genomics”) centers on seeing how diet influences quality articulation. Nutrigenomics is an examination field particularly which relies upon the new improvement of cutting-edge innovations that permit us to handle a lot of information identifying with quality variations. These alleged “-omics” advances genomics, proteomics, metabolomics, and transcriptomics, which permit us to recognize and gauge various kinds of atoms at the same time.

29.6 Developing genomic technologies for enhancing food crops security

Understanding the quality capacity and cooperation has the option to build up a connection between the living being’s genome and its aggregate. Various procedures that are broadly used to comprehend the quality/protein work incorporate RNA impedance (RNAi), mutagenesis, mass spectrometry, genome explanation, etc. The vast majority of the useful genomic explores are done on model types of plants/creatures/people since model life forms offer a savvy approach to follow the legacy of qualities through numerous ages in a moderately brief time frame. Similar genomics approaches can be additionally used to comprehend the greater genomes dependent on the information got from model creatures. Genomics is a broad methodology that has been designed to upscale the foreseeing facilities, communications of qualities, and function of genomics. The same depicted in previous research studies that stages of genome sequencing headway have been made possible to completely grouping countless plant genomes. A blast of quality grouping data has represented a significant test of distinguishing qualities and deciding their capacity. The genomics period presently took imperative changes in a practical genomics for handling of a few key inquiries concerning employed at various stages along with tissues explicitness. Be that as it may, coordination and investigation of the genomic information is the greatest test nowadays. Some of the online workers are dealing with the utilization of quality bioinformatics sequencing (Lohse, Lang, & Boyd, 2014). The HT sequence-nucleotide polymorphism sequencing exhibits intended in genome-wide affiliation of QTL considers (Chen et al., 2014). The practical utilization is one of the significant utilities for changes in the quality guideline (Mieulet, Diévar, Droc, Lanau, & Guiderdoni, 2013). A few converse hereditary qualities apparatuses, for example, mutagenesis in the transposons, T-DNA, impedance of RNAi and focusing on incited nearby sores in genomes (TILLING), and empower analysts to contemplate explicit qualities (Chen et al., 2014). The presentation of transposable labeling Ac–Ds elements in a maize framework extends extraordinary occasions to connect qualities through work with making portraying alleles of freak. Also, infection-prompted quality quieting has been considered a fast and practical useful examination instrument for species yield (Stratmann & Hind, 2011). Plowing (focusing on initiated nearby injuries in genomes) is another generally acknowledged opposite hereditary methodology that is at present being utilized to screen the populace for transformations in objective qualities. On the other hand, sequential examination of quality articulation serial analysis of gene expression (SAGE), massively parallel signature sequencing (MPSS), and microarrays are accessible during the plant harvesting for the profile of RNA (mRNA) concurrent expectation to follow the action of countless qualities. Notwithstanding, the previously revealed atomic techniques, biochemical apparatuses, for example, proteomics and metabolomics are likewise assuming a significant part to follow the quality profiling of protein and expression of metabolites (Gupta, Langridge, & Mir, 2010). The International Rice Functional Genomics Steering Committee held in 2020 expected toward the organization of rice-based practical genomics research. The major centers are the way to recognize the elements in the genome of rice quality for expanding creation toward focusing on food security complexity (Zhang et al., 2008). A forecast of quality capacity based on practical genomics choices to expand food creation and quality nutritional value in significant assortments of foods. Manipulation in the methodologies facilitates the distinguishing proof in the account of important qualities governing characteristic related to agronomics through a financial incentive (Singh, Gupta, & Rai, 2013). Through ceaseless advancement in the perspective of genomics devices, reproducers could build up modern assortments lenient associated with various kinds of stresses, including biotic and abiotic. As of late, in wheat assortments of Ug99 stem rustproof was created for TILLING utilization (Kim et al., 2014). Accessibility of new genomics coordinated stages long for improving new assortments dependable and productive.

29.7 Application of high-throughput genomics technologies in nutrition research

It decides the structures of all the protein crease families encoded by the qualities of living creatures. On the off chance that fruitful, this will permit the structures, all things considered, or quality items to be controlled by homology to proteins where the area overlay structure has been settled. Information on a protein’s structure and its homology to different proteins give bits of knowledge into the capacity of the protein and its parts inside organic frameworks. This information may permit us to adjust a protein’s movement with inhibitor atoms or activator particles and by hereditary

designing. All such potential outcomes depend on a comprehension of the protein's physical, compound, and mathematical properties, concluded from its atomic structure. Underlying genomics means fundamentally describing most protein arrangements by a proficient blend of trial and forecast. This point will be accomplished via a cautious choice of target proteins and their structure assurance by X-beam crystallography or NMR spectroscopy. There is an assortment of target determination plans, going from zeroing in on just novel folds to choosing all proteins in a model genome. A model-driven view necessitates that objectives be chosen to such an extent that the greater part of the excess arrangements can be demonstrated with helpful precision by near displaying. Indeed, even with underlying genomics, the structure of the vast majority of the proteins will be demonstrated, not dictated by test. As examined over, the exactness of near models and correspondingly the assortment of their applications decline strongly beneath the 30% grouping personality cutoff, fundamentally because of a fast expansion in arrangement mistakes. In this manner, we should decide protein structures so the vast majority of the excess groupings are identified within any event one known structure at higher than 30% arrangement character. It was as of late assessed that this cutoff requires at least 16,000 focuses to cover 90% of all protein area families, including those of film proteins. These 16,000 structures will permit the demonstrating of a particularly bigger number of proteins. For instance, New York Structural Genomics Research Consortium estimated the effect of its structures by recording the number and nature of the comparing models for perceptibly related proteins in the nonrepetitive grouping information base. Overall, 100 protein arrangements with no earlier primary portrayal could be demonstrated in any event at the overlay level. This huge influence of structure assurance by protein structure displaying outlines and legitimizes the reason for primary genomics. One more structure forecast will be added to primary genomics severally. Enormous scope all over again forecast can direct objective choice by zeroing in on trial structure assurance on proteins liable to embrace novel folds. One more strategy should likewise be valuable in supplementing near displaying techniques by building parts of proteins not present in format structures. What is more, new techniques enhanced by restrictions from cross-connecting or different trials can give models to proteins not promptly agreeable to X-beam crystallographic or NMR investigation. At long last, enormous scope all over again displaying may permit coarse structure-based bits of knowledge into protein capacity of countless proteins well ahead of time of tentatively decided structures.

References

- Agrawal, S., Devanur, N.R., & Li, L. (2016, June). An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory* (pp. 4–18). PMLR.
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.
- Gong, M., Li, Y., Wang, H., Liang, Y., Wu, J. Z., Zhou, J., & Dai, H. (2013). An advanced Ni–Fe layered double hydroxide electrocatalyst for water oxidation. *Journal of the American Chemical Society*, *135*(23), 8452–8455.
- Gu, H., Zhu, P., Jiao, Y., Meng, Y., & Chen, M. (2011). PRIN: a predicted rice interactome network. *BMC bioinformatics*, *12*(1), 1–13.
- Gupta, P. K., Langridge, P., & Mir, R. R. (2010). Marker-assisted wheat breeding: present status and future possibilities. *Molecular Breeding*, *26*(2), 145–161.
- Helmy, M., Tomita, M., & Ishihama, Y. (2011). OryzaPG-DB: rice proteome database based on shotgun proteogenomics. *BMC plant biology*, *11*(1), 1–9.
- Kim, S., Kim, D., Cho, S. W., Kim, J., & Kim, J. S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome research*, *24*(6), 1012–1019.
- Lohse, K. R., Lang, C. E., & Boyd, L. A. (2014). Is more better? Using metadata to explore dose–response relationships in stroke rehabilitation. *Stroke*, *45*(7), 2053–2058.
- Mieulet, D., Diévar, A., Droc, G., Lanau, N., & Guiderdoni, E. (2013). Reverse genetics in rice using Tos17. In *Plant Transposable Elements*, (pp. 205–221). Totowa, NJ: Humana Press.
- Que, S., Li, K., Chen, M., Wang, Y., Yang, Q., Zhang, W., & He, H. (2012). PhosphoRice: a meta-predictor of rice-specific phosphorylation sites. *Plant Methods*, *8*(1), 1–9.
- Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., Li, A., & Higuchi, K. (2013). Non-orthogonal multiple access (NOMA) for cellular future radio access. In *2013 IEEE 77th vehicular technology conference*, (pp. 1–5). VTC Spring.
- Singh, K. P., Gupta, S., & Rai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, *80*, 426–437.
- Stratmann, J. W., & Hind, S. R. (2011). Gene silencing goes viral and uncovers the private life of plants. *Entomologia experimentalis et applicata*, *140*(2), 91–102.
- The French-Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*, 463–467, CrossRefGoogle Scholar.

- Wang, Y., Ying, Q., Hu, J., & Zhang, H. (2014). Spatial and temporal variations of six criteria air pollutants in 31 provincial capital cities in China during 2013–2014. *Environment international*, *73*, 413–422.
- Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., & Williams, K. L. (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and genetic engineering reviews*, *13* (1), 19–50.
- Zhang, Y. H., Hu, M., Zhong, L. J., Wiedensohler, A., Liu, S. C., Andreae, M. O., & Fan, S. J. (2008). Regional integrated experiments on air quality over Pearl River Delta 2004 (PRIDE-PRD2004): Overview. *Atmospheric Environment*, *42*(25), 6157–6173.

Further reading

- Ainouche, M. L., & Jenczewski, E. (2010). Focus on polyploidy. *The New Phytologist*, *186*, 1–4, CrossRefGoogle Scholar.
- Allender, C. J., & King, G. J. (2010). Origins of the amphiploid species *Brassica napus* L. investigated by chloroplast and nuclear molecular markers. *BMC Plant Biology*, *10*, 54, CrossRefGoogle Scholar.
- Amborella Genome Project, A.G. (2013). The Amborella genome and the evolution of flowering plants. *Science (New York, N.Y.)*, *342*, 1241089, CrossRefGoogle Scholar.
- Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, *546*, 148–152, CrossRefGoogle Scholar.
- Bancroft, I., Morgan, C., Fraser, F., Higgins, J., Wells, R., Clissold, L., et al. (2011). Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nature Biotechnology*, *29*, 762–766, CrossRefGoogle Scholar.
- Bayer, P. E., Hurgobin, B., Golicz, A. A., Chan, C.-K. K., Yuan, Y., Lee, H., et al. (2017). Assembly and comparison of two closely related *Brassica napus* genomes. *Plant Biotechnology Journal*, *15*, 1602–1610, CrossRefGoogle Scholar.
- Beenakker, C. W. J. (2006). Specular andreev reflection in graphene. *Physical Review Letters*, *97*(6). Available from <https://doi.org/10.1103/PhysRevLett.97.067007>, 067,007, URL.
- Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., & Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 18724–18728, CrossRefGoogle Scholar.
- Bowers, J. E., Chapman, B. A., Rong, J., & Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, *422*, 433–438, CrossRefGoogle Scholar.
- Bus, A., Hecht, J., Huettel, B., Reinhardt, R., & Stich, B. (2012). High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *BMC Genomics*, *13*, 281, CrossRefGoogle Scholar.
- Cai, G., Yang, Q., Yi, B., Fan, C., Edwards, D., Batley, J., & Zhou, Y. (2014). A complex recombination pattern in the genome of allotetraploid *Brassica napus* as revealed by a high-density genetic map. *PLoS One*, *9*, e109910, CrossRefGoogle Scholar.
- Chadwick, R. (2004). Nutrigenomics, individualism and public health. *Proceedings of the Nutrition Society*, *63*(1), 161–166.
- Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., et al. (2014). Early allopolyploid evolution in the post-neolithic *Brassica napus* oilseed genome. *Science (New York, N.Y.)*, *345*, 950–953, CrossRefGoogle Scholar.
- Chen, X., Li, X., Zhang, B., Xu, J., Wu, Z., Wang, B., Li, H., Younas, M., Huang, L., Luo, Y., et al. (2013). Detection and genotyping of restriction fragment associated with pseudo sequence polymorphisms in polyploid in *Brassica napus* crops. *BMC Genomics*, *14*, 346, CrossRefGoogle Scholar.
- Clarke, W. E., Higgins, E. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., . . . Parkin, I. A. P. (2016). A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *129*, 1887–1899, CrossRefGoogle Scholar.
- Clarke, W. E., Parkin, I. A., Gajardo, H. A., Gerhardt, D. J., Higgins, E., Sidebottom, C., . . . Iniguez-Luy, F. L. (2013). Genomic DNA enrichment using sequence capture microarrays: A novel approach to discover sequence nucleotide polymorphisms (SNP) in *Brassica napus* L. *PLoS One*, *8*, e81992, CrossRefGoogle Scholar.
- Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews. Genetics*, *6*, 836–846, CrossRefGoogle Scholar.
- Dalton-Morgan, J., Hayward, A., Alamery, S., Tollenaere, R., Mason, A. S., Campbell, E., Patel, D., Lorenc, M. T., Yi, B., Long, Y., et al. (2014). A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Functional & Integrative Genomics*, *14*, 643–655, CrossRefGoogle Scholar.
- Dassanayake, M., Oh, D. H., Haas, J. S., Hernandez, A., Hong, H., & Ali, S. (2013). The genome of the extremophile crucifer *Thellungiella parvula*. *Nature Genetics*, *43*, 913–918, CrossRefGoogle Scholar.
- Delourme, R., Falentin, C., Fomeju, B., Boillot, M., Lassalle, G., André, I., Duarte, J., Gauthier, V., Lucante, N., Marty, A., et al. (2013). High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. *BMC Genomics*, *14*, 120, CrossRefGoogle Scholar.
- Dubcovsky, J., & Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science (New York, N.Y.)*, *316*, 1862–1866, CrossRefGoogle Scholar.
- Edwards, D., Batley, J., & Snowdon, R. J. (2013). Accessing complex crop genomes with next-generation sequencing. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *126*, 1–11, CrossRefGoogle Scholar.

- Gaeta, R., Pires, J., Iniguez-Luy, F., Leon, E., & Osborn, T. (2007). Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *The Plant Cell*, *19*, 3403–3417, CrossRefGoogle Scholar.
- Gaeta, R. T., & Pires, J. C. (2010). Homoeologous recombination in allopolyploids: The polyploid ratchet. *The New Phytologist*, *186*, 18–28, CrossRefGoogle Scholar.
- Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 1513–1518, CrossRefGoogle Scholar.
- Griffin, A. S., West, S. A., & Buckling, A. (2004). Cooperation and competition in pathogenic bacteria. *Nature*, *430*, 1024–1027.
- Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., & Williamson, R. J. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, *45*, 891–898, CrossRefGoogle Scholar.
- He, Z., Wang, L., Harper, A. L., Havlickova, L., Pradhan, A. K., Parkin, I. A. P., & Bancroft, I. (2017). Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnology Journal*, *15*, 594–604, CrossRefGoogle Scholar.
- Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., & Clark, R. M. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, *43*, 476–481, CrossRefGoogle Scholar.
- Huang, S., Deng, L., Guan, M., Li, J., Lu, K., Wang, H., . . . Hua, W. (2013). Identification of genome-wide single nucleotide polymorphisms in allopolyploid crop *Brassica napus*. *BMC Genomics*, *14*, 717, CrossRefGoogle Scholar.
- Jackson, S., & Chen, Z. J. (2010). Genomic and expression plasticity of polyploidy. *Current Opinion in Plant Biology*, *13*, 153–159, CrossRefGoogle Scholar.
- Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmockel, S. M., Li, B., Borm, T. J. A., Ohyanagi, H., Mineta, K., Michell, C. T., Saber, N., et al. (2017). The genome of *Chenopodium quinoa*. *Nature*, *542*, 307–312, CrossRefGoogle Scholar.
- Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C. S., et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*, 524–527, PubMedGoogle Scholar.
- Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., . . . Soltis, P. S. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*, 97–100, CrossRefGoogle Scholar.
- Johnston, J. S., Pepper, A. E., Hall, A. E., Chen, Z. J., Hodnett, G., Drabek, J., . . . Price, H. J. (2005). Evolution of genome size in Brassicaceae. *Annals of Botany*, *95*, 229–235, CrossRefGoogle Scholar.
- Kagale, S., Robinson, S. J., Nixon, J., Xiao, R., Huebert, T., Condie, J., Kessler, D., Clarke, W. E., Edger, P. P., Links, M. G., et al. (2014). Polyploid evolution of the Brassicaceae during the Cenozoic Era. *The Plant Cell*, *26*, 2777–2791, CrossRefGoogle Scholar.
- Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*, 1384–1395, CrossRefGoogle Scholar.
- Lampe, J. W., et al. (2013). Inter-individual differences in response to dietary intervention: Integrating omics platforms towards personalised dietary recommendations. *Proceedings of the Nutrition Society*, *72*(2), 207–218.
- Larkan, N. J., Lydiate, D. J., Parkin, I. A. P., Nelson, M. N., Epp, D. J., Cowling, W. A., . . . Borhan, M. H. (2013). The *Brassica napus* blackleg resistance gene LepR3 encodes a receptor-like protein triggered by the *Leptosphaeria maculans* effector AVR/LM1. *The New Phytologist*, *197*, 595–605, Google Scholar.
- Liu, C., Wang, J., Huang, T., Wang, F., Yuan, F., Cheng, X., . . . Liu, K. (2010). A missense mutation in the VHYNP motif of a DELLA protein causes a semi-dwarf mutant phenotype in *Brassica napus*. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *121*, 249–258, CrossRefGoogle Scholar.
- Liu, J., Hua, W., Hu, Z., Yang, H., Zhang, L., Li, R., . . . Wang, H. (2015). Natural variation in ARF18 gene simultaneously affects seed weight and silique length in polyploid rapeseed. *Proc Natl Acad Sci USA*, *112*, 5123–5132, CrossRefGoogle Scholar.
- Liu, L., Qu, C., Wittkop, B., Yi, B., Xiao, Y., He, Y., . . . Li, J. (2013). A high-density SNP map for accurate mapping of seed fibre QTL in *Brassica napus* L. *PLoS One*, *8*, e83052, CrossRefGoogle Scholar.
- Liu, L., Stein, A., Wittkop, B., Sarvari, P., Li, J., Yan, X., . . . Snowdon, R. J. (2012). A knockout mutation in the lignin biosynthesis gene CCR1 explains a major QTL for acid detergent lignin content in *Brassica napus* seeds. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *124*, 1573–1586, CrossRefGoogle Scholar.
- Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A., Zhao, M., Ma, J., Yu, J., Huang, S., et al. (2014). The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*, *5*, 3930, CrossRefGoogle Scholar.
- Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*, 18, Google Scholar.
- Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., et al. (2008). Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiology*, *148*, 1772–1781, CrossRefGoogle Scholar.
- Lysak, M. A., Koch, M. A., Pecinka, A., & Schubert, I. (2005). Chromosome triplication found across the tribe Brassicaceae. *Genome Research*, *15*, 516–525, CrossRefGoogle Scholar.
- Mason, A. S., Higgins, E. E., Stein, A., Werner, C., Batley, J., Parkin, I. A. P., & Snowdon, R. J. (2017). A user guide to the Brassica 60 K Illumina Infinium™ SNP genotyping array. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, *130*, 621–633, CrossRefGoogle Scholar.

- Masterson, J. (1994). Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science (New York, N.Y.)*, 264, 421–424, CrossRefGoogle Scholar.
- Mathers, J. C. (2017). Nutrigenomics in the modern era. *Proceedings of the Nutrition Society*, 76(3), 265–275.
- Mead, M. N. (2007). Nutrigenomics – The Genome-Food interface. *Environmental Health Perspectives*, 115(12), 582–589.
- Neumann, G. A., & Mazarico, E. (2009). Planetary science. Seeing the missing half. *Science (New York, N.Y.)*, 323, 885–887.
- Ordovas, J. M., et al. (2018). Personalised nutrition and health. *The British Medical Journal*, 361, bmj.k2173.
- Osborn, T. C., Butrulle, D. V., Sharpe, A. G., Pickering, K. J., Parkin, I. A. P., Parker, J. S., & Lydiat, D. J. (2003). Detection and effects of a homeologous reciprocal transposition in *Brassica napus*. *Genetics*, 165, 1569–1577, PubMedPubMedCentralGoogle Scholar.
- Ott, J., Kamatani, Y., & Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews. Genetics*, 12, 465–474, CrossRefGoogle Scholar.
- Parkin, I. A., Gulden, S. M., Sharpe, A. G., Lukens, L., Trick, M., Osborn, T. C., & Lydiat, D. J. (2005). Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics*, 171, 765–781, CrossRefGoogle Scholar.
- Parkin, I. A. P., Sharpe, A. G., Keith, D. J., & Lydiat, D. J. (1995). Identification of the A and C genomes of amphidiploid *Brassica napus* (oilseed rape). *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 38, 1122–1131, CrossRefGoogle Scholar.
- Petronis, A. (2010). Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, 465, 721–727.
- Pires, J. C., Zhao, J., Schranz, M. E., Leon, E. J., Quijada, P. A., Lukens, L. N., & Osborn, T. C. (2004). Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biological Journal of the Linnean Society*, 82, 675–688, CrossRefGoogle Scholar.
- Raman, H., Dalton-Morgan, J., Diffey, S., Raman, R., Alamery, S., Edwards, D., & Batley, J. (2014). SNP markers-based map construction and genome-wide linkage analysis in *Brassica napus*. *Plant Biotechnology Journal*, 12, 851–860, CrossRefGoogle Scholar.
- Rana, D., Boogaart, T., O'Neill, C. M., Hynes, L., Bent, E., Macpherson, L., . . . Bancroft, I. (2004). Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. *The Plant Journal: for Cell and Molecular Biology*, 40, 725–733, CrossRefGoogle Scholar.
- Rousseau-Gueutin, M., Morice, J., Coriton, O., Huteau, V., Trotoux, G., Nègre, S., Falentin, C., Deniot, G., Gilet, M., Eber, F., et al. (2016). The impact of open pollination on the structural evolutionary dynamics, meiotic behavior and fertility of resynthesized allotetraploid *Brassica napus* L. G3: Genes. *Genomes Genetics*, 7, 705–717, Google Scholar.
- Rygulla, W., Friedt, W., Seyis, F., Lühs, W., Eynck, C., Tiedemann, A. V., & Snowdon, R. J. (2007a). Combination of resistance to *Verticillium longisporum* from zero erucic acid *Brassica oleracea* and oilseed *Brassica rapa* genotypes in resynthesized rapeseed (*Brassica napus*) lines. *Plant Breeding*, 126, 596–602, CrossRefGoogle Scholar.
- Rygulla, W., Snowdon, R. J., Eynck, C., Koopmann, B., von Tiedemann, A., Lühs, W., & Friedt, W. (2007b). Broadening the genetic basis of *Verticillium longisporum* resistance in *Brassica napus* by interspecific hybridization. *Phytopathology*, 97, 1391–1396, CrossRefGoogle Scholar.
- Schnable, J. C., Springer, N. M., & Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 4069–4074, CrossRefGoogle Scholar.
- Sharpe, A. G., Parkin, I. A. P., Keith, D. J., & Lydiat, D. J. (1995). frequent nonreciprocal translocations in the amphidiploid genome of oilseed rape (*Brassica napus*). *Genome/National Research Council Canada = Genome/Conseil National de Recherches Canada*, 38, 1112–1121, CrossRefGoogle Scholar.
- Slotte, T., Hazzouri, K. M., Ågren, J. A., Koenig, D., Maumus, F., & Guo, Y. L. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics*, 45, 831–835, CrossRefGoogle Scholar.
- Snowdon, R. J., Abbadi, A., Kox, T., Schmutzer, T., & Leckband, G. (2015). Heterotic haplotype capture: Precision breeding for hybrid performance. *Trends in Plant Science*, 20, 410–413, CrossRefGoogle Scholar.
- Snowdon, R. J., & Iniguez Luy, F. L. (2012). Potential to improve oilseed rape and canola breeding in the genomics era. *Plant Breeding*, 131, 351–360, CrossRefGoogle Scholar.
- Soltis, D., & Soltis, P. (2003). The role of phylogenetics in comparative genetics. *Plant Physiology*, 132, 1790–1800, CrossRefGoogle Scholar.
- Song, K., Lu, P., Tang, K., & Osborn, T. C. (1995). Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 92, 7719–7723, CrossRefGoogle Scholar.
- Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S. V., Obermeier, C., . . . Snowdon, R. J. (2017). Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. *Plant Biotechnology Journal*, 15, 1478–1489, CrossRefGoogle Scholar.
- Sun, F., Fan, G., Hu, Q., Zhou, Y., Guan, M., Tong, C., Li, J., Du, D., Qi, C., Jiang, L., et al. (2017). The high-quality genome of *Brassica napus* cultivar ‘ZS11’ reveals the introgression history in semi-winter morphotype. *The Plant Journal: for Cell and Molecular Biology*, 92, 452–468, CrossRefGoogle Scholar.
- Szadkowski, E., Eber, F., Huteau, V., Lodé, M., Huneau, C., Belcram, H., Coriton, O., Manzanares-Dauleux, M. J., Delourme, R., King, G. J., et al. (2010). The first meiosis of resynthesized *Brassica napus*, a genome blender. *The New Phytologist*, 186, 102–112, CrossRefGoogle Scholar.
- The World Health Organisation. (2014). *Global Status Report on non-communicable diseases*. Geneva: WHO Press.
- Trick, M., Long, Y., Meng, J., & Bancroft, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal*, 7, 334–346, CrossRefGoogle Scholar.
- Udall, J. A., Quijada, P. A., & Osborn, T. C. (2005). Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of *Brassica napus* L. *Genetics*, 169, 967–979, CrossRefGoogle Scholar.

- Van de Peer Y. (2011). A mystery unveiled. *Genome Biology*, 12, 113, CrossRefGoogle Scholar.
- Vision, T. J., Brown, D. G., & Tanksley, S. D. (2000). The origins of genomic duplications in Arabidopsis. *Science (New York, N.Y.)*, 290, 2114–2117, CrossRefGoogle Scholar.
- Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., & Liu, S. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, 43, 1035–1039, CrossRefGoogle Scholar.
- Werner, S., Diederichsen, E., Frauen, M., Schondelmaier, J., & Jung, C. (2007). Genetic mapping of clubroot resistance genes in oilseed rape. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 116, 363, CrossRefGoogle Scholar.
- Xia, S., Wang, Z., Zhang, H., Hu, K., Zhang, Z., Qin, M., Dun, X., Yi, B., Wen, J., Ma, C., et al. (2016). Altered transcription and neofunctionalization of duplicated genes rescue the harmful effects of a chimeric gene in *Brassica napus*. *The Plant Cell*, 28, 2060–2078, CrossRefGoogle Scholar.
- Xiong, Z., Gaeta, R. T., & Pires, J. C. (2011). Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 7908–7913, CrossRefGoogle Scholar.
- Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., Hu, Z., Chen, S., Pental, D., Ju, Y., et al. (2016). The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nature Genetics*, 48, 1225–1232, CrossRefGoogle Scholar.
- Yi, B., Zeng, F., Lei, S., Chen, Y., Yao, X., Zhu, Y., Wen, J., Shen, J., Ma, C., Tu, J., et al. (2010). Two duplicate CYP704B1-homologous genes BnMs1 and BnMs2 are required for pollen exine formation and tapetal development in *Brassica napus*. *The Plant Journal*, 63, 925–938, CrossRefGoogle Scholar.
- Zhang, D., Hua, Y., Wang, X., Zhao, H., Shi, L., & Xu, F. (2014). A high-density genetic map identifies a novel major QTL for boron efficiency in oilseed rape (*Brassica napus* L.). *PLoS One*, 9, e112089, CrossRefGoogle Scholar.
- Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics (Oxford, England)*, 29, 2669–2677, CrossRefGoogle Scholar.

This page intentionally left blank

Bioinformatics approach for whole transcriptomics-based marker prediction in agricultural crops

Habeeb Shaik Mohideen¹, Archit Gupta¹ and Sewali Ghosh²

¹Bioinformatics and Entomoinformatics Lab, Department of Genetic Engineering, School of Bioengineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India, ²Department of Zoology and Advanced Biotechnology, Guru Nanak College, Chennai, Tamil Nadu, India

30.1 Introduction to transcriptomics

30.1.1 Transcriptome

The collection of all the coding information present in a cell called transcriptome describes the functional identity of the genome. It is a very complex entity to study, and its composition is easily affected by several environmental, physical, chemical, and other stresses. The impact of these factors on the transcriptome gives us a better idea about the gene regulatory events in a cell (Lowe, Shirley, Bleackley, Dolan, & Shafee, 2017). Transcriptomics is the study of transcriptome data that includes collecting raw data, data processing and analysis, obtaining meaningful information, and applying the knowledge gained for the transformative betterment.

There is a lot of information to be obtained through transcriptomics, which can be utilized for more applications. These include gene expression cycles of a cell in a tissue, protein expression changes in response to a stimulus, and how a drug can alter gene expression. Such studies can be used for finding treatments and cures for diseases, help identifying drought-resistant genes, counter drug-resistant microbes, and many other similar applications.

30.2 Markers

A marker is a unique reference sequence and acts as the identity of a gene or species condition, disease state, species identification, and quantification. Over the years, markers are at the center of many techniques like mapping, trait association, diversity and evolutionary analysis, and marker-assisted breeding for crop variety improvement (Hayward, Tollenaere, Dalton-Morgan, & Batley, 2015). In addition, protein markers are already being used in molecular diagnostics and understand protein expression in a cell. Markers can be broadly classified into four types: phenotypic, biochemical, cytological, and molecular.

30.2.1 Phenotypic markers

Phenotypic markers usually are not the gene of interest themselves but are a sequence near to them, usually tightly linked to the gene of interest, so the presence of the marker can be used to confirm the presence of the gene of interest, or vice versa. In this case the gene of interest usually plays a role in the phenotype of an organism (Prasenjit, Anirudha, Gautam, Jaya, & Sonam, 2017). Though these markers are readily available and their effects are visible, they lack a diversity of traits, suffer from Mendelian limitations, are easily influenced by environmental factors leading to false results, and using them in breeding experiments is labor-intensive and time taking.

30.2.2 Biochemical markers

Biochemical markers are functional markers which are fully characterized from the gene sequence to the protein function. However, they can be identified through highly accurate assays and are known to increase selection efficiency for traits with low inheritability (Lande & Thompson, 1990). Nevertheless, very few genes have been functionally (Andersen & Lübberstedt, 2003) characterized, and only a few of them can actually be used as markers and may also require expensive technology and assays for identification.

30.2.3 Cytological markers

Cytological markers correspond to variations in chromosomes, which can be a variation in chromosomal number, size, order, position, or banding pattern. They also help distinguish between heterochromatin and euchromatin. These include the G-banding (Giemsa staining), Q-banding (Quinacrine hydrochloride staining), and R-banding (inverted G-banding) (Nadeem et al., 2017). Thus they are beneficial in genetic mapping and the identification of linkage groups. But they are very limited in count and variety, and the results produced require expertise to understand.

30.2.4 Molecular markers

Molecular markers are DNA-based ones. They provide much higher throughput and unparalleled precision than other markers. Any DNA sequence stretch can be used as a marker as long as it is unique and has no off-target sites. They are proven to be accurate, easy to reproduce the results generated, and are plentiful. However, the expensive nature and its preciseness turn out to be the limitations of these categories of markers.

30.3 Markers in plants

Markers play an essential role in the selective breeding of plants (marker-assisted selection). They are used to improve traits that are considered desirable and to remove undesirable traits from plants (Jiang, 2013). In addition, transcriptomics plays an important role, as cDNA is made from transcriptome and used to obtain the exonic region of the DNA, to be used as markers (Adhikari et al., 2017). Some of the techniques used for the identification of these markers are listed in Table 30.1.

TABLE 30.1 Techniques used for identification of markers.

RFLP—restriction fragment length polymorphism (Botstein, White, Skolnick, & Davis, 1980)
AFLP—amplified-fragment length polymorphism (Vos et al., 1995)
CAPS or PCR—RFLP—cleaved amplified polymorphic sequences or polymerase chain reaction—restriction fragment length polymorphism (Konieczny & Ausubel, 1993)
DArT—diversity array technology (Konieczny & Ausubel, 1993)
SSCP—single-strand conformation polymorphism (Hayashi & Yandell, 1993; Hayashi, 1992; Paran & Michelmore, 1993)
F-SSCP—fluorescence-based PCR-SSCP (Hayashi, 1992; Makino et al., 1992)
ISSR—inter-simple sequence repeats (Meyer, Mitchell, Freedman, & Vilgalys, 1993; Zietkiewicz, Rafalski, & Labuda, 1994)
MP-PCR—microsatellite-primed polymerase chain reaction (Weising, Nybom, Pfenninger, Wolff, & Meyer, 1994)
Microsatellites or SSRs (simple sequence repeats) or STRs (short tandem repeats) or SSLPs (simple sequence length polymorphisms) (Hamada & Kakunaga, 1982; Litt & Luty, 1989; Tautz, Trick, & Dover, 1986)
Minisatellites or DAMD (directed amplification of minisatellite DNA) or VNTRs (variable number of tandem repeats) (Jeffreys, Wilson, & Thein, 1985; Jeffreys, Neumann, & Wilson, 1990)
RAPD—randomly amplified polymorphic DNA (Williams, Pande, Nair, Mohan, & Bennett, 1991)
RAMP—randomly amplified microsatellite polymorphisms (Williams et al., 1991)
SCAR—sequence characterized amplified regions (McDermott et al., 1994; Paran & Michelmore, 1993)
SRAP—sequence-related amplified polymorphism (Li & Quiros, 2001)
SNPs—single-nucleotide polymorphisms (Brookes, 1999; Cho et al., 1999)
SPAR—single-primer amplification reaction (Gupta, Chyi, Romero-Severson, & Owen, 1994)
SCoT—start codon targeted polymorphism (Gupta et al., 1994)
TRAP—targeted region amplification polymorphism (Jinguo & Vick, 2003; Xiaohua, Deyuan, & Zhenhui, 2004)
Sequencing (Behjati & Tarpey, 2013; Mardis, 2008; Mardis, 2013; Maxam & Gilbert, 1977; Pareek, Smoczynski, & Tretyn, 2011)

30.4 Expressed sequence tags and simple sequence repeats

ESTs stand for expressed sequence tags. They are short polynucleotides made from a single RNA molecule (Lowe et al., 2017) and are generated during RNA sequencing (RNA-Seq). RNA is converted into cDNA by the use of reverse transcriptase (RTase) (Goff, 1990) and then sequenced (Marra, Hillier, & Waterston, 1998). For using ESTs, we do not need any prior knowledge of the organism from which they are made, and thus they can be used with a wider variety of organisms. In addition, EST libraries are used to provide information on sequence for microarrays; for example, a GeneChip for barley (*Hordeum vulgare*) was compiled from about 350,000 ESTs (Close et al., 2004).

SSRs stand for simple sequence repeats, also known as microsatellites. SSR markers are made from EST datasets (Feng et al., 2016) and are used as markers for germplasm analysis, among other uses (Powell et al., 1996). When compared to other molecular marker techniques like RAPD (Bardakci, 2001; Hadrys, Balick, & Schierwater, 1992) (random amplified polymorphic DNA) and AFLP (Bachem et al., 1996; Mueller & Wolfenbarger, 1999; Vos et al., 1995) (amplified fragment length polymorphism), SSRs proved to be the most reproducible technique (Jones et al., 1997). They are considered one of the best markers due to their immense application in identifying genotypes, construction of genomic maps, and marker–trait association (Kujur et al., 2013).

30.5 Tools for generating transcriptomic data

30.5.1 Serial analysis of gene expression technology

SAGE stands for serial analysis of gene expression. We can use this technology to obtain the global gene expression profile of a cell or a tissue or use it to find genes that are differentially expressed between a pair or a group of cells (Velculescu, Zhang, Vogelstein, & Kinzler, 1995; Yamamoto, Wakatsuki, Hada, & Ryo, 2001). In this method, short reads of 8–12 bp of cDNA are produced and then assembled to obtain sequence data (Velculescu et al., 1995; Yamamoto et al., 2001). In some variations of SAGE, like the longSAGE (Wei et al., 2004) or Robust-Long-SAGE (Gowda, Jantasuriyarat, Dean, & Wang, 2004) methods, reads can be 17 bp or longer (Hu & Polyak, 2006). The amount of reads for a particular sequence is directly proportional to the level of gene expressions (Velculescu, Vogelstein, & Kinzler, 2000).

30.5.2 Microarrays

Microarrays allow the analysis of hundreds or even thousands of parameters, all at once (DeRisi et al., 1996; Lockhart et al., 1996; Schena, Shalon, Davis, & Brown, 1995; Schena et al., 1996; Schena, 1996). This allows the technology to extract high throughput gene expression data and, with high precision, better than previous methods and technologies (Grunstein & Hogness, 1975; Southern, 2000). In this method the mRNA is converted into cDNA and then hybridized with labeled probes, all done on a chip. The binding of the probe to the target gives a signal, which is recorded by a computer. The signals between multiple samples are analyzed to obtain differential gene expression profiles (Kaliyappan, Palanisamy, Govindarajan, & Duraiyan, 2012; Murphy, 2002; Stears, Martinsky, & Schena, 2003).

30.5.3 RNA sequencing

RNA sequencing or RNA-Seq is a sequencing technique based on the NGS (next-generation sequencing) principles. They are used to study the transcriptome, which is a very dynamic entity (Marguerat & Bähler, 2010). Since then, it has been the gold standard for RNA sequencing, to predict and annotate genes, and to study transcriptional and posttranscriptional gene regulation (Haas & Zody, 2010; Marguerat & Bähler, 2010). RNAseq has quickly replaced chip-based technologies like microarrays as the preferred technique for studying the differential expression of genes in different environments due to easier use and more reliable and reproducible results (Haas & Zody, 2010; Hiremath et al., 2011; Tarazona, García-Alcalde, Dopazo, Ferrer, & Conesa, 2011).

Basic protocol followed for marker identification (Mudalkar, Golla, Ghatty, & Reddy, 2014) are as follows:

1. Plant source: multiple samples of the plant source with the desired traits are collected.
2. RNA extraction: total RNA is extracted from the samples.
3. cDNA library construction: cDNA library is made using RTase and RNaseH enzymes.
4. cDNA sequencing: sequencing can be done using any method.
5. Sequence assembly: the FASTQ reads generated through sequencing are quality-checked and assembled.

6. Data analysis: the best transcripts are chosen, with the highest number of exons, longest open reading frames, and highest confidence score.
7. Gene and pathway annotation: the transcripts are searched in existing databases using BLASTx, and the annotations were given for the sequence with the highest similarity.
8. Sequence similarity: we try to find similar sequences in related organism's transcriptome or proteome.
9. Marker prediction: if all criteria are met and the sequence is unique, they can be used as markers and help predict traits (Fig. 30.1).

30.6 Why transcriptomic markers?

Genomics has been used in crops for quite some time. Genomic markers have the potential to improve crops significantly (Malmberg et al., 2018). Many tools have been created, but only a small number of them have seen proper use, because plant varieties have significantly different genome sizes and ploidy. In addition, they have a high diversity in single-nucleotide polymorphism (SNP) frequency (Kaliyappan et al., 2012; Murphy, 2002; Stears et al., 2003). Due to this, we need to develop different markers for every variety of even the same crop, which makes the process unfeasible and very expensive. On the contrary, GBS-t (genotyping-by-sequencing through transcriptomics) is emerging to be a cost-effective, high throughput, and most importantly, broadly applicable system (Fu & Yang, 2017; Kim et al., 2016; Malmberg et al., 2018). Generally, systems like the GBS-t and RNA-Seq have been used for plants for which a reference genome is available to align the reads and annotate them (Fu & Yang, 2017; Kim et al., 2016), novel pipelines like the GB-eaSy (Wickland, Battu, Hudson, Diers, & Hudson, 2017), UGbS-Flex (Qi et al., 2018), and more have been developed for transcriptomic sequencing when such genomic data are not available.

Plants display the highest variance in genomic size (number of chromosomes can range from 2 to more than 600), ploidy (1 to more than 20), and genetic diversity (haplophase genome differs by over 2500 folds, that to just in angiosperms) (Bennett, 2008; Mohammadi & Prasanna, 2003). This means the DNA sequence of each plant is highly variable, leading to a problem when using genetic markers developed for one plant in a different plant (Dong, Liu, Yu, Wang, & Zhou, 2012; Ouborg, Piquot, & Van Groenendael, 1999). Since we are dealing with the transcriptome, any changes due to silent/synonymous SNPs are eliminated, leading to a broader scope of marker development. Thus SNPs contribute to a significant fraction of genomic polymorphism seen in plants. There have been several types of research confirming this, some of which we will look into next.

In a study conducted with the maize plant (*Zea mays*), scientists (Azodi, Pardo, VanBuren, de Los Campos, & Shiu, 2020) developed genetic and transcriptome-based markers. They ran benchmarks for accuracy of trait predictions for both and found out that transcriptome data–based markers give us a better link between traits (Seo et al., 2016). They can better capture data, which is not easily possible when using DNA sequence–based markers. The reason for that could be that a change in DNA sequence does not necessarily mean a difference in protein (and thus the trait) due to the degenerate nature of the genetic code (Seo et al., 2016). According to the study, genetic markers could identify only 1 out of 14 benchmark flowering-time genes, while transcriptome-based markers identified 5 out of the 14 (Azodi et al., 2020). It gives transcriptome- or proteome-based markers an edge over DNA markers in trait prediction. It proves that the edge transcriptome- or proteome-based markers have over DNA markers in trait prediction.

Genetic markers are particular, sometimes to a fault. They cannot be used for species other than those designed for (Eckert, Samis, & Lougheed, 2008; Elshire et al., 2011; Selkoe & Toonen, 2006). But in the case of transcriptome markers, a study (Stokes et al., 2010) found that markers made for a dicot crop can be used for monocot crops as well. They looked at markers made for *Arabidopsis thaliana* (dicot), corresponding to increased vegetative biomass and increased yield of consumable grains. They then used those markers to see if they could also help in predicting or

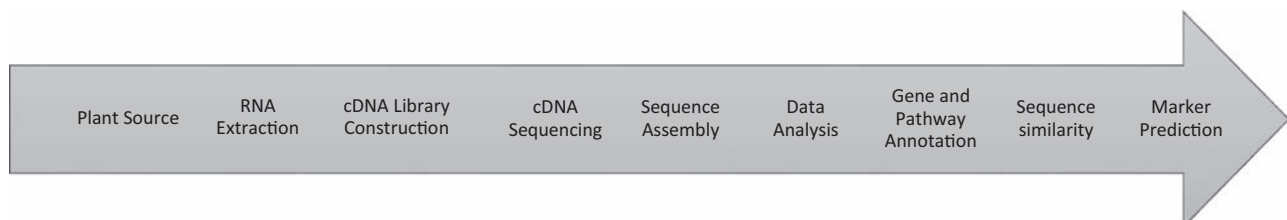


FIGURE 30.1 A simple pipeline for marker identification from the RNA-Seq raw data. *RNA-Seq*, RNA sequencing.

selecting maize (monocot) plants. They found that the markers were valid for maize as well. The study indicates the superior nature of transcriptomic markers over genetic markers.

In yet another study a selection of crops with different ploidy levels, genetic polymorphism, and breeding methods was taken. This included perennial ryegrass, which is an outbreeding diploid ($2n = 2x = 14$) forage grass; Phalaris, which is an outbreeding allotetraploid ($2n = 4x = 28$) forage grass; lentils, an inbreeding diploid ($2n = 2x = 14$) legume; and canola, a partially outbreeding allotetraploid ($2n = 4x = 38$) oilseed. The transcriptomes of samples from each of the selected crops were sequenced using the following GBS-t pipeline (Malmberg et al., 2018):

The obtained sequences were compared to existing data on these plants and cross-examined within the species samples. Both outbreeding and inbreeding crops were selected at allotetraploid and diploid levels to validate the method. The pipeline generated 89,738 to 231,977 SNPs, with good genomic coverage (c.3 million sequence reads for every sample). For perennial ryegrass, 83 samples with high diversity were selected, resulting in 139,772 SNP loci in a total of 11,787 reference transcriptome contigs (Fig. 30.2).

For lentils, 182 samples were taken, presenting 38 ancestral genotypes. A total of 231,977 SNPs were identified corresponding to 30,573 contigs, out of which 85% (25,897) were placed uniformly in the 7 chromosomes of the lentil plant. For Phalaris, 285 samples were taken from a breeding program, representing key genotypes. A total of 89,738 sites were selected. For canola, 575 different lines were taken. They consisted of spring lines from Australia and winter lines from discrete geographical locations. As a result, 76,270 SNPs were identified, which were shared between both types.

The identified transcriptomic markers for each crop were usable with other varieties of the crops. The study (Malmberg et al., 2018) then concludes the GBS-t analysis as a widely applicable system, which is relevant for various crops, given the choice of sequence analysis software is appropriate. It also proved to be a relatively more straightforward and automated system, with superior cost-effectiveness and data usability.

30.7 How are markers developed/selected?

Sesame (*Sesamum indicum* L.) is one of the oldest and most crucial seed crops for oil production. The seed contains approximately 45%–58% oil. Illumina paired-end technology was used to sequence sesame transcriptomes. The acquired reads were assembled to obtain 86,222 unigenes with a mean size of 629 bp. Out of these, 46,584 had similarities with proteins from Swiss-Prot and NCBI NR databases. Upon investigation in the KEGG database, 22,003 of these unigenes were placed in 119 pathways. On searching through BLASTx, 15,460 of the total unigenes showed homology to *Arabidopsis* genes (TAIR database, version 10). Thus a total of 7702 unigenes were made into EST–SSR markers.

The rubber tree (*Hevea brasiliensis* Muell. Arg.) is a significant tree for the commercial production of rubber. The main component of the tree is the bark. As important as it is, there is a severe lack of transcriptome data for the bark.

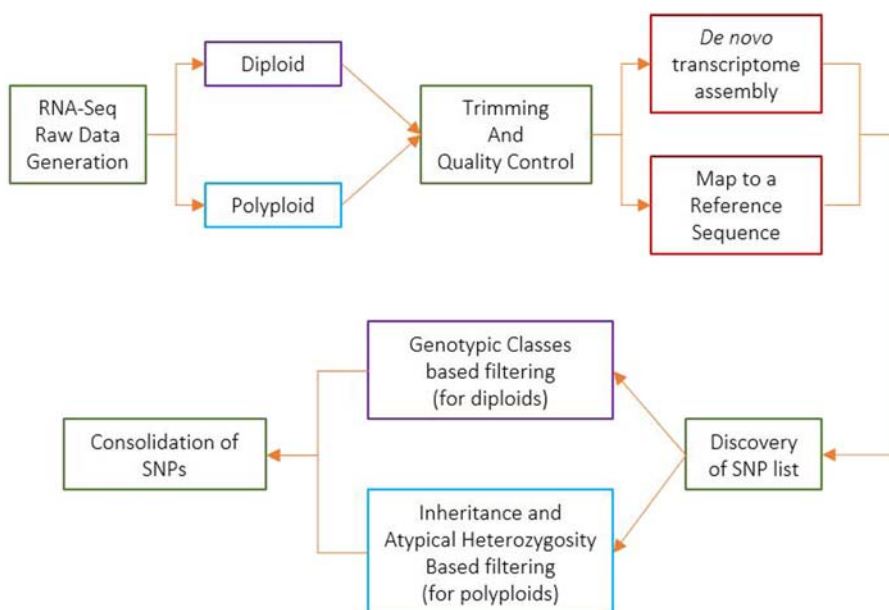


FIGURE 30.2 An overview of the GBS-t pipeline. GBS-t, Genotyping-by-sequencing through transcriptomics.

A study was done in 2012, which also generated transcriptomic markers for trait prediction (Li, Deng, Qin, Liu, & Men, 2012). They used Illumina paired-end sequencing to generate over 30 million reads. On performing de novo assembly, they identified 22,756 unigenes with a mean length of 485 base pairs. They were then run against the Swiss-Prot and NCBI NR databases, and they got 12,558 and 16,520 hits, respectively. They also identified 39,257 ESTs of SSR (EST–SSRs). Out of these, 110 markers were selected to predict higher bark metabolism for the trees.

30.8 What has been done

In the 1990s, some of the earliest works in the transcriptome were done (Piétu et al., 1999; Velculescu et al., 1997). Sanger sequencing (Stranneheim & Lundeberg, 2012; Valencia, Pervaiz, Husami, Qian, & Zhang, 2013) principles, SAGE technology was developed to sequence the transcriptome (Velculescu et al., 1995). Soon after, microarray technology became popular and played a crucial role in transcriptomic studies (Ben-Dor, Shamir, & Yakhini, 1999; Girke et al., 2000; Reymond, Weber, Damond, & Farmer, 2000; Schena et al., 1995; Schena et al., 1996). Microarray became a better technology for the characterization of gene expression due to its higher throughput (Kavsan et al., 2007; Kim, 2003).

Chickpea (*Cicer arietinum* L.) is a significant crop in Africa and Asia, where water scarcity and droughts are a perpetual threat. There has been improved crop yield, but not much, due to high stress on the crops, both biotic and abiotic. To improve, researchers used NGS methods Illumina/Solexa and Roche/454 to sequence the transcriptome. They compared the chickpea transcriptome with the transcriptome of *Medicago truncatula*, a model legume plant closely associated with chickpea and a degree of resistance to drought. The researchers identified 728 SSRs, 387 orthologous sequences, conserved 495 SNPs, and 2088 intronic markers (Hiremath et al., 2011). As the similarity was found to be with a drought-resistant plant, these similar regions must contribute toward the trait. Therefore these in silico predicted markers can be used for marker-assisted breeding to produce chickpea varieties suitable to be grown in these regions, which will reduce the drought-induced stress on the plants and result in a better yield of the crop.

European ash (*Fraxinus excelsior*) is a significant source of wood in Europe and is traditionally used as the material of choice for various tools. But a good portion of the trees in cultivation is susceptible to a fungal pathogen, the *Hymenoscyphus fraxineus*, causing the ash dieback disease. The disease is chronic and is characterized by crown dieback and leaf loss. To reduce the loss of crops a group of researchers identified transcriptomic markers (Harper et al., 2016), which were closely associated with deterioration in the canopy of the infected trees.

To identify these markers, they sequenced the transcriptome using 100 bp Illumina HiSeq reads, which were aligned to 41,521 known unigenes, and CLC Transcript Discovery plugin (Schäuser, 2019) gene models were constructed. mRNA-Seq reads were mapped from a group of 182 diverse trees, and 470,494 SNPs were identified. After eliminating SNPs with low allele frequency and low transcript abundance mean, 32,441 GEMs (gene expression markers) were deemed suitable for use as markers.

Using these markers, in a group of unrelated trees, they successfully predicted the tree samples that were more susceptible to canopy damage. Thus eliminating such plants at an early stage would result in much higher crop yield, and reduce the loss of crops, thus increasing the supply and making it more profitable.

30.9 Future prospects

With the massive amounts of transcriptomic data we have and the more significant amount generated in this RNA-Seq era, the future for the improvement of crops is brighter than ever (Egan, Schlueter, & Spooner, 2012; Flanagan & Jones, 2019; Hamilton & Robin, 2012; Zhuang, Zhang, Hou, Wang, & Xiong, 2014). Some highlights of what we can achieve in the future with a wider variety of crops and with better accuracy and precision:

- improved drought tolerance for crops (Cattivelli et al., 2008; Degenkolbe et al., 2013; Frova, Villa, Sari-Gorla, Krajewski, & Di Fonzo, 1999; Mir, Zaman-Allah, Sreenivasulu, Trethowan, & Varshney, 2012; Sprenger et al., 2018; Tuinstra, Ejeta, & Goldsbrough, 1998);
- selection of better crops through marker-assisted breeding (Collard & Mackill, 2008; Collard, Jahufer, Brouwer, & Pang, 2005; Jiang, 2016; Ribaut & Hoisington, 1998);
- improving crop yield (Gouda et al., 2020; Ribaut, Jiang, Gonzalez-de-Leon, Edmeades, & Hoisington, 1997; Steele et al., 2013; Stuber, Edwards, & Wendel, 1987; Stuber, Polacco, & Senior, 1999);
- better resistance to plant pathogens and diseases (Bent et al., 1994; Melchinger, 1990; Miah et al., 2013; Yang et al., 2012; Young, 1996);

- better herbicide tolerance by selecting plants with genes for herbicide tolerance (Grover et al., 2020; Marshall et al., 1992; Milligan, Daly, Parry, Lazzeri, & Jepson, 2001; Sari-Gorla et al., 1997);
- improved salt tolerance (Ashraf & Foolad, 2013; Ashraf, 2009; Foolad & Jones, 1993; Ganie, Wani, Henry, & Hensel, 2021; Luo et al., 2017);
- better production and higher yield of medicinal drug (Graham et al., 2010; Jelodar, Bhatt, Profile, Mohamed, & Chan, 2014; Zhang et al., 2018);
- reduced susceptibility to pests (Haley, Afanador, & Kelly, 1994; Lefebvre & Chèvre, 1995; Levi et al., 2013; Mori et al., 2011);
- improved tolerance to being submerged in cases of floods, etc. Lefebvre and Chèvre (1995; Haley et al., 1994; Levi et al., 2013; Mori et al., 2011);
- higher production and yield of secondary plant metabolites (Ahmad, Shahzad, Sharma, & Parveen, 2018; Chavan et al., 2014; Piątczak, Kuźma, Sitarek, & Wysokińska, 2015);
- better quality food grains (Hari et al., 2011; Heffner, Jannink, Iwata, Souza, & Sorrells, 2011; Jairin et al., 2009; Luo et al., 2014; Sukumaran et al., 2012); and
- improved quality of bark for wood extraction (Fady et al., 2003; Grattapaglia, Bertolucci, Penchel, & Sederoff, 1996; Lenz et al., 2017; Thavamanikumar et al., 2014; Williams & Neale, 1992).

Other than these, there are many more distinct prospects for using transcriptomic markers in agricultural crop improvement.

References

- Adhikari, S., Saha, S., Biswas, A., Rana, T. S., Bandyopadhyay, T. K., & Ghosh, P. (2017). *Nuclear*, 60, 283.
- Ahmad, Z., Shahzad, A., Sharma, S., & Parveen, S. (2018). *Plant Cell, Tissue and Organ Culture*, 132, 497.
- Andersen, J. R., & Lübberstedt, T. (2003). *Functional markers in plants* (Vol. 8, pp. 554–560). Elsevier Ltd.
- Ashraf, M. (2009). Biotechnological approach of improving plant salt tolerance using antioxidants as markers. *Biotechnology Advances*, 27, 84–93.
- Ashraf, M., & Foolad, M. R. (2013). In R. Tuberosa (Ed.), *Crop breeding for salt tolerance in the era of molecular markers and marker-assisted selection* (Vol. 132, pp. 10–20). .
- Azodi, C. B., Pardo, J., VanBuren, R., de Los Campos, G., & Shiu, S. H. (2020). *The Plant Cell*, 32, 139.
- Bachem, C. W. B., van der Hoeven, R. S., de Bruijn, S. M., Vreugdenhil, D., Zabeau, M., & Visser, R. G. F. (1996). *The Plant Journal: For Cell and Molecular Biology*, 9, 745.
- Bardakci, F. (2001). Random amplified polymorphic DNA (RAPD) markers. *The Scientific and Technological Research Council of Turkey*.
- Behjati, S., & Tarpey, P. S. (2013). *Archives of Disease in Childhood: Education and Practice Edition*, 98, 236.
- Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). *Journal of Computational Biology*, 281–297.
- Bennett, M. D. (2008). *The New Phytologist*, 106, 177.
- Bent, A. F., Kunkel, B. N., Dahlbeck, D., Brown, K. L., Schmidt, R., Giraudat, J., . . . Staskawicz, B. J. (1994). *Science*, 265, 1856.
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). *Construction of a genetic linkage map in man using restriction fragment length polymorphisms* (Vol. 32, pp. 314–331). Elsevier.
- Brookes, A. J. (1999). *The essence of SNPs* (Vol. 234, pp. 177–186). Elsevier.
- Cattivelli, L., Rizza, F., Badeck, F. W., Mazzucotelli, E., Mastrangelo, A. M., Francia, E., . . . Stanca, A. M. (2008). *Drought tolerance improvement in crop plants: An integrated view from breeding to genomics. Field Crops Research*, 105, 1–14.
- Chavan, J. J., Gaikwad, N. B., Umdale, S. D., Kshirsagar, P. R., Bhat, K. V., & Yadav, S. R. (2014). *Plant Growth Regulation*, 72, 1.
- Cho, R. J., Mindrinos, M., Richards, D. R., Sapolsky, R. J., Anderson, M., Drenkard, E., . . . Oefner, P. J. (1999). *Nature Genetics*, 23, 203.
- Close, T. J., Wanamaker, S. I., Caldo, R. A., Turner, S. M., Ashlock, D. A., Dickerson, J. A., . . . Wise, R. P. (2004). *Plant Physiology*, 134, 960.
- Collard, B. C. Y., & Mackill, D. J. (2008). *Marker-assisted selection: An approach for precision plant breeding in the twenty-first century* (Vol. 363, pp. 557–572). Royal Society.
- Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). *An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica*, 142, 169–196.
- DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., . . . Trent, J. M. (1996). *Nature Genetics*, 14, 457.
- Degenkolbe, T., Do, P. T., Kopka, J., Zuther, E., Hinch, D. K., & Köhl, K. I. (2013). *PLoS One*, 8, e63637.
- Dong, W., Liu, J., Yu, J., Wang, L., & Zhou, S. (2012). *PLoS One*, 7, e35071.
- Eckert, C. G., Samis, K. E., & Lougheed, S. C. (2008). *Genetic variation across species' geographical ranges: The central-marginal hypothesis and beyond. Molecular Ecology*, 17, 1170–1188.
- Egan, A. N., Schlueter, J., & Spooner, D. M. (2012). *American Journal of Botany*, 99, 175.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). *PLoS One*, 6, e19379.
- Fady, B., Ducci, F., Aleta, N., Becquey, J., Diaz Vazquez, R., Fernandez Lopez, F., . . . Rumpf, H. (2003). *New Forest*, 25, 211.
- Feng, S., He, R., Lu, J., Jiang, M., Shen, X., Jiang, Y., . . . Wang, H. (2016). *Frontiers in Genetics*, 7, 113.

- Flanagan, S. P., & Jones, A. G. (2019). *Molecular Ecology*, 28, 544.
- Foolad, M. R., & Jones, R. A. (1993). *TAG Theoretical and Applied Genetics*, 87, 184.
- Frova, C., Villa, M., Sari-Gorla, M., Krajewski, P., & Di Fonzo, N. (1999). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 99, 280.
- Fu, Y. B., & Yang, M. H. (2017). *Methods in Molecular Biology* (pp. 169–187). Humana Press Inc.
- Ganie, S. A., Wani, S. H., Henry, R., & Hensel, G. (2021). *Improving rice salt tolerance by precision breeding in a new era* (Vol. 60, p. 101996). Elsevier Ltd.
- Gerke, T., Todd, J., Ruuska, S., White, J., Benning, C., & Ohlrogge, J. (2000). *Plant Physiology*, 124, 1570.
- Goff, S. P. (1990). *Journal of Acquired Immune Deficiency Syndromes (1999)*, 3, 817.
- Gouda, G., Gupta, M. K., Donde, R., Mohapatra, T., Vadde, R., & Behera, L. (2020). *Marker-assisted selection for grain number and yield-related traits of rice (Oryza sativa L.)* (Vol. 26, pp. 885–898). Springer.
- Gowda, M., Jantasuriyarat, C., Dean, R. A., & Wang, G. L. (2004). *Plant Physiology*, 134, 890.
- Graham, L. A., Besser, K., Blumer, S., Branigan, C. A., Czechowski, T., Elias, L., . . . Bowles, D. (2010). *Science*, 327, 328.
- Grattapaglia, D., Bertolucci, F. L. G., Penchel, R., & Sederoff, R. R. (1996). *Genetics*, 144.
- Grover, N., Kumar, A., Yadav, A. K., Gopala Krishnan, S., Ellur, R. K., Bhowmick, P. K., . . . Singh, A. K. (2020). *Rice*, 13, 68.
- Grunstein, M., & Hogness, D. S. (1975). *Proceedings of the National Academy of Sciences of the United States of America.*, 72, 3961.
- Gupta, M., Chyi, Y. S., Romero-Severson, J., & Owen, J. L. (1994). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 89, 998.
- Haas, B. J., & Zody, M. C. (2010). *Advancing RNA-Seq analysis*, 28, 421–423.
- Hadrys, H., Balick, M., & Schierwater, B. (1992). *Molecular Ecology*, 1, 55.
- Haley, S. D., Afanador, L., & Kelly, J. D. (1994). *Crop Science*, 34, 1061.
- Hamada, H., & Kakunaga, T. (1982). *Nature*, 298, 396.
- Hamilton, J. P., & Robin Buell, C. (2012). *Advances in plant genome sequencing*, 70, 177–190.
- Hari, Y., Srinivasarao, K., Viraktamath, B. C., Hariprasad, A. S., Laha, G. S., Ahmed, M. I., . . . Sundaram, R. M. (2011). *Plant Breeding*, 130, 608.
- Harper, A. L., McKinney, L. V., Nielsen, L. R., Havlickova, L., Li, Y., Trick, M., . . . Bancroft, I. (2016). *Scientific Reports.*, 6, 1.
- Hayashi, K. (1992). *PCR-SSCP: A method for detection of mutations* (Vol. 9, pp. 73–79). Elsevier.
- Hayashi, K., & Yandell, D. W. (1993). *Human Mutation*, 2, 338.
- Hayward, A. C., Tollenaere, R., Dalton-Morgan, J., & Batley, J. (2015). *Methods in Molecular Biology*, 1245, 13.
- Heffner, E. L., Jannink, J. L., Iwata, H., Souza, E., & Sorrells, M. E. (2011). *Crop Science*, 51, 2597.
- Hiremath, P. J., Farmer, A., Cannon, S. B., Woodward, J., Kudapa, H., Tuteja, R., . . . Varshney, R. K. (2011). *Plant Biotechnology Journal*, 9, 922.
- Hu, M., & Polyak, K. (2006). *Nature Protocols*, 1, 1743.
- Jairin, J., Teangdeerith, S., Leelagud, P., Kothcharerk, J., Sansen, K., Yi, M., . . . Toojinda, T. (2009). *Field Crops Research*, 110, 263.
- Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). *Nature*, 314, 67.
- Jeffreys, A. J., Neumann, R., & Wilson, V. (1990). *Cell*, 60, 473.
- Jelodar, N. B., Bhatt, A., Profile, S., Mohamed, K., & Chan, L. K. (2014). *Journal of Medicinal Plant Research*.
- Jiang, G.-L. (2013). *Plant breeding from laboratories to fields*. InTech.
- Jiang, G. L. (2016). *Advances in plant breeding strategies: Breeding, biotechnology and molecular tools* (pp. 431–472). Cham: Springer International Publishing.
- Jinguo, H. U., & Vick, B. A. (2003). *Plant Molecular Biology Reporter*, 21, 289.
- Jones, C. J., Edwards, K. J., Castaglione, S., Winfield, M. O., Sala, F., Van De Wiel, C., . . . Karp, A. (1997). *Molecular Breeding*, 3, 381.
- Kaliyappan, K., Palanisamy, M., Govindarajan, R., & Duraiyan, J. (2012). *Journal of Pharmacy and Bioallied Sciences*, 4, 310.
- Kavsan, V. M., Dmitrenko, V. V., Shostak, K. O., Bukreieva, T. V., Vitak, N. Y., Simirenko, O. E., . . . Zozulya, Y. A. (2007). *Cytology and Genetics.*, 41, 36.
- Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L. S., & Paterson, A. H. (2016). *Application of genotyping by sequencing technology to a variety of crop breeding programs* (Vol. 242, pp. 14–22). Ireland Ltd: Elsevier.
- Kim, H. L. (2003). *Experimental & Molecular Medicine*, 35, 460.
- Konieczny, A., & Ausubel, F. M. (1993). *The Plant Journal: for Cell and Molecular Biology*, 4, 403.
- Kujur, A., Bajaj, D., Saxena, M. S., Tripathi, S., Upadhyaya, H. D., Gowda, C. L. L., . . . Parida, S. K. (2013). *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 20, 355.
- Lande, R., & Thompson, R. (1990). *Genetics*, 124.
- Lefebvre, V., & Chèvre, A. M. (1995). *Agronomie*, 15, 3.
- Lenz, P. R. N., Beaulieu, J., Mansfield, S. D., Clément, S., Despoints, M., & Bousquet, J. (2017). *BMC Genomics*, 18, 335.
- Levi, A., Thies, J. A., Wechter, W. P., Harrison, H. F., Simmons, A. M., Reddy, U. K., . . . Fei, Z. (2013). *Genetic Resources and Crop Evolution*, 60, 427.
- Li, D., Deng, Z., Qin, B., Liu, X., & Men, Z. (2012). *BMC Genomics*, 13, 1.
- Li, G., & Quiros, C. F. (2001). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 103, 455.
- Litt, M., & Luty, J. A. (1989). *American Journal of Human Genetics*, 44, 397.
- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., . . . Brown, E. L. (1996). *Nature Biotechnology*, 14, 1675.

- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). *PLoS Computational Biology*, 13, e1005457.
- Luo, M., Zhao, Y., Zhang, R., Xing, J., Duan, M., Li, J., ... Zhao, J. (2017). *BMC Plant Biology*, 17, 140.
- Luo, Y., Zakaria, S., Basyah, B., Ma, T., Li, Z., Yang, J., & Yin, Z. (2014). *Rice*, 7, 33.
- Makino, A., Sakashita, H., Hidema, J., Mae, T., Ojima, K., & Osmond, B. (1992). *Plant Physiology*, 100, 1737.
- Malmberg, M. M., Pembleton, L. W., Baillie, R. C., Drayton, M. C., Sudheesh, S., Kaur, S., ... Cogan, N. O. I. (2018). *Plant Biotechnology Journal*, 16, 877.
- Mardis, E. R. (2008). *Annual Review of Genomics and Human Genetics*, 9, 387.
- Mardis, E. R. (2013). *Annual Review of Analytical Chemistry*, 6, 287.
- Marguerat, S., & Bähler, J. (2010). *RNA-seq: From technology to biology* (Vol. 67, pp. 569–579). Springer.
- Marra, M. A., Hillier, L., & Waterston, R. H. (1998). *Trends in Genetics: TIG*, 14, 4.
- Marshall, L. C., Somers, D. A., Dotray, P. D., Gengenbach, B. G., Wyse, D. L., & Gronwald, J. W. (1992). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 83, 435.
- Maxam, A. M., & Gilbert, W. (1977). *Proceedings of the National Academy of Sciences of the United States of America*, 74, 560.
- McDermott, J. M., Brandle, U., Dutly, F., Haemmerli, U. A., Keller, S., Müller, K. E., & Wolfe, M. S. (1994). *Genetic variation in powdery mildew of barley: Development of RAPD, SCAR, and VNTR markers*, 84, 1316–1321.
- Melchinger, A. E. (1990). *Use of molecular markers in breeding for oligogenic disease resistance. Plant Breeding*, 104, 1–19.
- Meyer, W., Mitchell, T. G., Freedman, E. Z., & Vilgalys, R. (1993). *Journal of Clinical Microbiology*, 31.
- Miah, G., Raffii, M. Y., Ismail, M. R., Puteh, A. B., Rahim, H. A., Islam, N. K., & Latif, M. A. (2013). *A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. International Journal of Molecular Sciences*, 14, 22499–22528.
- Milligan, A. S., Daly, A., Parry, M. A. J., Lazzeri, P. A., & Jepson, I. (2001). *Molecular Breeding*, 7, 301.
- Mir, R. R., Zaman-Allah, M., Sreenivasulu, N., Trethowan, R., & Varshney, R. K. (2012). *Integrated genomics, physiology and breeding approaches for improving drought tolerance in crops* (Vol. 125, pp. 625–645). Springer Verlag.
- Mohammadi, S. A., & Prasanna, B. M. (2003). *Crop Science*, 43, 1235.
- Mori, K., Sakamoto, Y., Mukojima, N., Tamiya, S., Nakao, T., Ishii, T., & Hosaka, K. (2011). *Euphytica*, 180, 347.
- Mudalkar, S., Golla, R., Ghatty, S., & Reddy, A. R. (2014). *Plant Molecular Biology*, 84, 159.
- Mueller, U. G., & Wolfenbarger, L. L. R. (1999). *AFLP genotyping and fingerprinting* (Vol. 14, pp. 389–394). Elsevier Ltd.
- Murphy, D. (2002). *Gene expression studies using microarrays: Principles, problems, and prospects* (Vol. 26, pp. 256–270). American Physiological Society.
- Nadeem, M. A., Amjad Nawaz, M., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., ... Shehzad Baloch, F. (2017). *International Journal of Agriculture Environment and Biotechnology*.
- Ouborg, N. J., Piquot, Y., & Van Groenendael, J. M. (1999). *Population genetics, molecular markers and the study of dispersal in plants*, 87, 551–568.
- Paran, I., & Michelmore, R. W. (1993). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 85, 985.
- Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). *Sequencing technologies and genome sequencing* (Vol. 52, pp. 413–435). Springer.
- Piétu, G., Mariage-Samson, R., Fayein, N. A., Matingou, C., Eveno, E., Houllgatte, R., ... Auffray, C. (1999). *Genome Research*, 9, 195.
- Piątczak, E., Kuźma, L., Sitarek, P., & Wysokińska, H. (2015). *Plant Cell, Tissue and Organ Culture*, 120, 539.
- Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S., & Rafalski, A. (1996). *The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis* (Vol. 2, pp. 225–238). The Netherlands: Springer.
- Prasenjit, D., Anirudha, S. K., Gautam, V., Jaya, B., & Sonam, M. (2017). *Marker and Its Application in Crop Improvement*.
- Qi, P., Gimode, D., Saha, D., Schröder, S., Chakraborty, D., Wang, X., ... Devos, K. M. (2018). *BMC Plant Biology*, 18.
- Reymond, P., Weber, H., Damond, M., & Farmer, E. E. (2000). *The Plant Cell*, 12, 707.
- Ribaut, J. M., & Hoisington, D. (1998). *Trends in Plant Science*, 3, 236.
- Ribaut, J. M., Jiang, C., Gonzalez-de-Leon, D., Edmeades, G. O., & Hoisington, D. A. (1997). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 94, 887.
- Sari-Gorla, M., Krajewski, P., Binelli, G., Frova, C., Taramino, G., & Villa, M. (1997). *Molecular Breeding*, 3, 481.
- Schauser, L. (2019). De novo transcriptome assembly using QIAGEN CLC Genomics Workbench. <https://digitalinsights.qiagen.com/news/blog/discovery/de-novo-transcript-assembly-using-clc-genomics-workbench/> (Accessed on 01 March 2022).
- Schena, M. (1996). *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology*, 18, 427.
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). *Science*, 270, 467.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., & Davis, R. W. (1996). *Proceedings of the National Academy of Sciences of the United States of America*, 93, 10614.
- Selkoe, K. A., & Toonen, R. J. (2006). *Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers*, 9, 615–629.
- Seo, M., Kim, K., Yoon, J., Jeong, J. Y., Lee, H. J., Cho, S., & Kim, H. (2016). *Scientific Reports*, 6, 24375.
- Southern, E. M. (2000). *Blotting at 25*, 25, 585–588.
- Sprenger, H., Erban, A., Seddig, S., Rudack, K., Thalhammer, A., Le, M. Q., ... Hinch, D. K. (2018). *Plant Biotechnology Journal*, 16, 939.
- Stears, R. L., Martinsky, T., & Schena, M. (2003). *Trends in Microarray Analysis*, 9, 140–145.
- Steele, K. A., Price, A. H., Witcombe, J. R., Shrestha, R., Singh, B. N., Gibbons, J. M., & Virk, D. S. (2013). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 126, 101.

- Stokes, D., Fraser, F., Morgan, C., O'Neill, C. M., Dreos, R., Magusin, A., ... Bancroft, I. (2010). *Molecular Breeding*, 26, 91.
- Stranneheim, H., & Lundeberg, J. (2012). *Biotechnology Journal*, 7, 1063.
- Stuber, C. W., Edwards, M. D., & Wendel, J. F. (1987). *Crop Science*, 27, 639.
- Stuber, C. W., Polacco, M., & Senior, M. L. (1999). *Crop Science*, 1571–1583.
- Sukumaran, S., Xiang, W., Bean, S. R., Pedersen, J. F., Kresovich, S., Tuinstra, M. R., ... Yu, J. (2012). *Plant Genome*, 5, plantgenome2012.07.0016.
- Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). *Genome Research*, 21, 2213.
- Tautz, D., Trick, M., & Dover, G. A. (1986). *Nature*, 322, 652.
- Thavamanikumar, S., McManus, L. J., Ades, P. K., Bossinger, G., Stackpole, D. J., Kerr, R., ... Tibbits, J. F. G. (2014). *Tree Genetics and Genomes*, 10, 1661.
- Tuinstra, M. R., Ejeta, G., & Goldsbrough, P. (1998). *Crop Science*, 38, 835.
- Valencia, C. A., Pervaiz, M. A., Husami, A., Qian, Y., & Zhang, K. (2013). *Next generation sequencing technologies in medical genetics* (pp. 3–11). New York, NY: Springer.
- Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). *Science*, 270, 484.
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., ... Kinzler, K. W. (1997). *Cell*, 88, 243.
- Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2000). *Analysing uncharted transcriptomes with SAGE. Trends in Genetics*, 16, 423–425.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Van De Lee, T., Hornes, M., ... Zabeau, M. (1995). *Nucleic Acids Research*, 23, 4407.
- Wei, C. L., Ng, P., Chiu, K. P., Wong, C. H., Ang, C. C., Lipovich, L., ... Ruan, Y. (2004). *Proceedings of the National Academy of Sciences of the United States of America*, 101, 11701.
- Weising, K., Nybom, H., Pfenninger, M., Wolff, K., & Meyer, W. (1994). *DNA Fingerprinting in Plants and Fungi* (1st ed.). CRC Press.
- Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). *BMC Bioinformatics*, 18.
- Williams, C. G., & Neale, D. B. (1992). *Canadian Journal of Forest Research. Journal Canadien de la Recherche Forestiere*, 22, 1009.
- Williams, M. N. V., Pande, N., Nair, S., Mohan, M., & Bennett, J. (1991). *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, 82, 489.
- Xiaohua, D., Deyuan, W., & Zhenhui, G. (2004). *Molecular Plant Breeding = Fen zi zhi wu yu Zhong*, 2, 740.
- Yamamoto, M., Wakatsuki, T., Hada, A., & Ryo, A. (2001). *Journal of Immunological Methods*, 250, 45.
- Yang, H., Tao, Y., Zheng, Z., Li, C., Sweetingham, M. W., & Howieson, J. G. (2012). *BMC Genomics*, 13, 318.
- Young, N. D. (1996). *QTL mapping and quantitative disease resistance in plants. Annual Review of Phytopathology*, 34, 479–501.
- Zhang, T., Bai, G., Han, Y., Xu, J., Gong, S., Li, Y., ... Liu, C. (2018). *Phytomedicine: International Journal of Phytotherapy and Phytopharmacology*, 44, 204.
- Zhuang, J., Zhang, J., Hou, X. L., Wang, F., & Xiong, A. S. (2014). *Critical Reviews in Plant Sciences*, 33, 225.
- Zietkiewicz, E., Rafalski, A., & Labuda, D. (1994). *Genomics*, 20, 176.

Computational approaches toward single-nucleotide polymorphism discovery and its applications in plant breeding

Dileep Kumar¹, Ranjana Gautam¹, Veda P. Pandey¹, Anurag Yadav², Upendra N. Dwivedi^{1,3}, Rumana Ahmad⁴ and Kusum Yadav¹

¹Department of Biochemistry, University of Lucknow, Lucknow, Uttar Pradesh, India, ²College of Basic Sciences and Humanities, Sardarkrushinagar Agricultural University Dantiwada, Palanpur, Gujarat, India, ³Institute for Development of Advanced Computing, ONGC Center for Advanced Studies, University of Lucknow, Lucknow, Uttar Pradesh, India, ⁴Department of Biochemistry, Era University, Lucknow, Uttar Pradesh, India

31.1 Introduction

In plant breeding the extent of genetic diversity is the prerequisite for developing/improving a new crop variety. Conventional plant breeding approach requires more than 12 years for the development of a crop variety; this long duration of time can be reduced by utilizing marker-assisted plant breeding that involves molecular marker techniques. In marker-assisted plant breeding the desired phenotypic trait is selected by identification of molecular markers which is derived from same region of genome where trait controlling gene is located (Jiang, 2013). Marker-based approaches have been extensively applied in crop improvement programs (Brookes, 1999; Hirschhorn & Daly, 2005; Rebbeck, Spitz, & Wu, 2004) such as random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), microsatellites or simple sequence repeats (SSRs), and single-nucleotide polymorphism (SNP) (Arif et al., 2010). Depending on detection method and throughput, molecular markers are divided into (1) low-throughput, hybridization-based markers such as restriction fragment length polymorphisms (RFLPs) (Botstein, White, Skolnick, & Davis, 1980); (2) medium-throughput, polymerase-chain reaction (PCR)-based markers that include RAPD (Welsh & McClelland, 1990), AFLP (Vos et al., 1995), SSRs (Jacob et al., 1991); (3) high-throughput (HTP) sequence-based markers, namely, SNPs (Wang et al., 1998). The use of these markers associated with crop yield has been applied to various crops such as rice (*Oryza sativa*) (Mackill, Nguyen, & Zhang, 1999), corn (*Zea mays*) (Ortiz, 2010), wheat (*Triticum aestivum*) (Suwarno, Pixley, Palacios-Rojas, Kaepler, & Babu, 2015), and tomatoes (*Lycopersicon esculentum*) (Landjeva, Korzun, & Börner, 2007).

In the late 1980s, RFLPs were the most popular molecular markers widely used in plant molecular genetics because they were reproducible and codominant (Lander & Botsteins, 1989). However, RFLP detection was an expensive, labor- and time-consuming unautomated process, which made this marker eventually obsolete. The invention of PCR technology and its application for the rapid detection of polymorphisms led to the new-generation PCR-based markers at the beginning of the 1990s. RAPD, AFLP, and SSR markers are such PCR-based markers that have been used by the research community in various plant systems. RAPDs are dominant markers that detect polymorphic loci in various genome regions (Williams, Kubelik, Livak, Rafalski, & Tingey, 1990). However, they are anonymous with a very low reproducibility level due to the nonspecific binding of short, random primers. Although AFLPs are also considered anonymous, their reproducibility and sensitivity level is higher owing to the longer +1 and +3 selective primers and the presence of discriminatory nucleotides at the 3' end of each primer. Consequently, AFLP markers are still popular in molecular genetics research in crops with little to zero reference genome sequence availability (Zhang, Guo, Liu, Tang, & Chen, 2011). However, AFLP

markers are not widely used in molecular breeding due to the lengthy and laborious detection method. On the other hand, the SSR markers are considered “markers of choice” for plant genome study (Powell, Machray, & Provan, 1996), as they are free from drawbacks of the abovementioned DNA markers. SSRs are not anonymous, highly reproducible, with high allelic polymorphism, and amenable to automation. Despite the detection cost remaining high, SSR markers had pervaded all plant molecular genetics and breeding areas in the late 1990s and the beginning of the 21st century.

SNPs comprise the most extensive set of sequence variants in most organisms (Kruglyak, 1997). First discovered in the human genome, SNPs emerged as the new-generation molecular markers, which have been proved to be universal and the most abundant forms of genetic variation among individuals of the same species. SNP refers to individual nucleotide base difference between two DNA sequences. These markers are abundant, ubiquitous, and amenable to high- and ultra-HTP automation (Foster et al., 2010). SNPs are classified according to nucleotide substitution as either transition (C/T or G/A) or transversions (C/G, A/T, C/A, or T/G). Although the presence of multiallelic SNPs is not exceptional, the SNPs are usually biallelic (two alternative bases occur) and require a minimum of 1% frequency in the population (Wang et al., 1998). As a nucleotide base is the smallest unit of inheritance, SNPs provide the ultimate form of the molecular genetic marker and the potential number of such markers is enormous in even closely related genotypes within a given species (Rafalski, 2002). The SNPs may aid in changing the genomic sequence, either in the coding (exons), intergenic, or noncoding (introns) region (Ahmad, Valentovic, & Rankin, 2018; Erwin et al., 2014). Being binary or codominant status, they can efficiently discriminate between homozygous and heterozygous alleles. SNPs are being utilized for biallelic mapping in diploid genomes and this has been established means for the creation of SNP-based maps to virtually any organism (Cho et al., 1999). The availability of various bioinformatics tools to retrieve and analyze big data (coming from genomics and transcriptomics) has discovered SNPs more efficient and fast (Kim, Kang, & Kim, 2020; Xu et al., 2017). Accordingly, SNP markers have been extensively utilized for different plant genetics and plant breeding applications for crop improvement in the last decade (Weckwerth, Ghatak, Bellaire, Chaturvedi, & Varshney, 2020).

Thus this chapter describes various computational approaches for SNP discovery in plants. Various databases and tools for SNP analysis, SNP genotyping, and their applications in plant breeding and crop improvement have also been discussed.

31.2 Single-nucleotide polymorphism discovery

Next-generation sequencing (NGS) is extensively used for DNA-sequencing, transcriptome sequencing, disease mapping, quantifying expression levels through RNA-sequencing, and population genetic studies (Metzker, 2010). With the advent of ever-increasing throughput in NGS, SNP mining in plants did not remain limited hence; de novo and reference-based SNP mining are now feasible for numerous plant species. NGS-derived SNPs are reported in humans (Altshuler et al., 2000), *Drosophila* (Berger et al., 2001), wheat (Allen et al., 2011; Trebbi et al., 2011), eggplant (Barchi et al., 2011), rice (Feltus et al., 2004; McNally et al., 2009; Yamamoto et al., 2010), *Arabidopsis* (Jander et al., 2002; Zhang & Borevitz, 2009), barley (Close et al., 2009; Waugh, Jannink, Muehlbauer, & Ramsay, 2009) sorghum (Nelson et al., 2011), cotton (Byers, Harker, Yourstone, Maughan, & Udall, 2012), common beans (Cortés, Chavarro, & Blair, 2011), soybean (Hyten et al., 2010), potato (Hamilton et al., 2011), flax (Fu & Peterson, 2012), *Aegilops tauschii* (You et al., 2011), alfalfa (Han et al., 2011), oat (Oliver et al., 2011), and maize (Jones et al., 2009), to name a few. SNP mining using NGS is readily accomplished in small plant genomes for which useful reference genomes are available such as rice and *Arabidopsis* (Ossowski et al., 2008; Yamamoto et al., 2010). However, SNP mining in plants without a reference genome sequence requires NGS data and, therefore, in such cases, several challenges persist. For instance, wheat (Allen et al., 2011; Trebbi et al., 2011), barley (Close et al., 2009; Waugh et al., 2009), oat (Oliver et al., 2011), beans (Cortés et al., 2011), and many others crops require SNP discovery via NGS.

SNP mining could be performed via two types of strategies: reference sequence strategy (Fig. 31.1) and de novo sequence strategy (Fig. 31.2). SNP mining from various NGS data encompasses the following steps: (1) grouping sequence reads according to their sequence similarity to identify reads covering the same part of the genome or having the same transcript origin, (2) aligning the reads, and (3) identifying and classifying sequence variants as potential polymorphisms.

31.2.1 Reference-based single-nucleotide polymorphism mining

The DNA and cDNA data generated from fully sequenced species can be used as the reference genome and reference transcriptome. Genome sequences of humans, animals, microbial species, and many plant species are examples of the ever-increasing reference groups, increasing research utilization for reference-based SNP mining. The HTP NGS

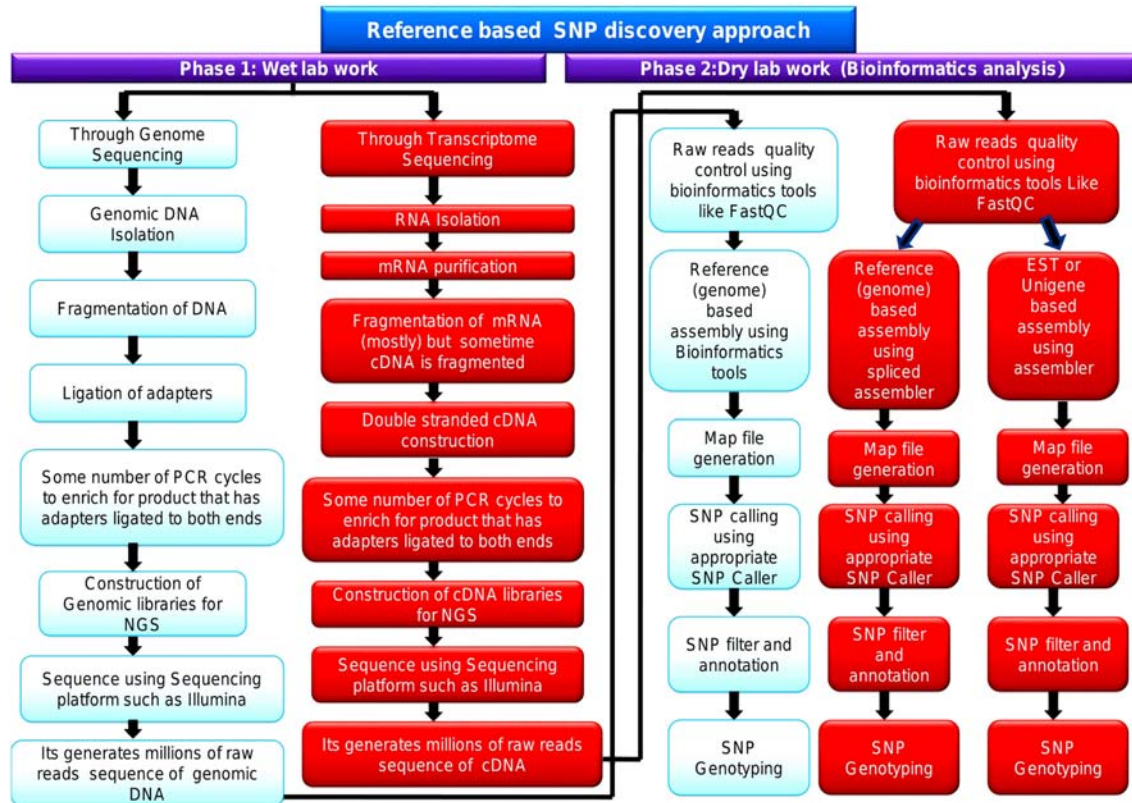


FIGURE 31.1 Workflow of SNP discovery and thereby its analysis through the reference-based strategy. This is the generalized workflow for computational SNP discovery and can be applied for SNP discovery in any plant that has deposited a reference genome/unigene/transcriptome database. *SNP*, Single-nucleotide polymorphisms.

platform's availability alleviated the slow sequencing process and exponentially increased data for the reference group. The reference sets can also be applied to a sequence that maps to a partially completed genome. When the reference sequence data of a species are available, a homology search tool must map the new sequence reads to the reference set. A global alignment tool or local alignment tool could be used. For example, the basic local sequence alignment tool (BLAST), sequence search and alignment by hashing algorithm (SSAHA) can perform this task (Ning, Cox, & Mullikin, 2001). Recently, several tools of sequence alignment such as Harvest (Treangen, Ondov, Koren, & Phillippy, 2014), TopHat 2 (Kim et al., 2016), Chain Cleaner (Suarez, Langer, Ladde, & Hiller, 2017), Kart (Lin & Hsu, 2017), and MUMmer4 (Marçais et al., 2018), are used for alignment of sequences. Another group of reference data can be generated from the PCR product, where primers are designed for a specified sequence region. Special tools such as Short Oligonucleotide Alignment Program (SOAP) (Li, Ruan, & Durbin, 2008), Mapping and Assembly with Qualities (MAQ) are used for mapping the reference data (Li, Li, Kristiansen, & Wang, 2008).

When transcriptome data are involved, it is easiest to map the data against a unigene set, resulting in an ungapped alignment. When such a dataset is not available, it can be mapped to genomic data using a spliced alignment tools like BLAT (Kent et al., 2002) or Spidey (Wheelan, Church, & Ostell, 2001), STAR (Dobin & Gingeras, 2016), ASGAL (Denti et al., 2018) SplicedFamAlign (Jammali, Aguilar, Kuitche, & Ouangraoua, 2019). The newly mapped data are aligned with the reference sequence. Pairwise or multiple alignments can evaluate base constitution on each position and the consequent SNP identification. Software tools such as Phrap (<http://www.phrap.org>) and CAP3 (Huang & Madan, 1999) are widely used for assembling the sequences to contigs. Multiple reads represent the sequence variants at each position. More sequence reads available in a species represent a specific genomic region with increased chances of finding a polymorphism. A sequence variant (allele) can also be distinguished from a sequencing error when multiple reads confirm it. The higher the number of reads per allele, the higher is the probability of it being a true polymorphism. The generalized workflow for SNP discovery based on reference sequence is shown in Fig. 31.1. As depicted in Fig. 31.1, SNP mining can be roughly divided into two phases, that is, wet lab work and dry lab work (bioinformatics analysis). The two phases are elaborated in the following sections.

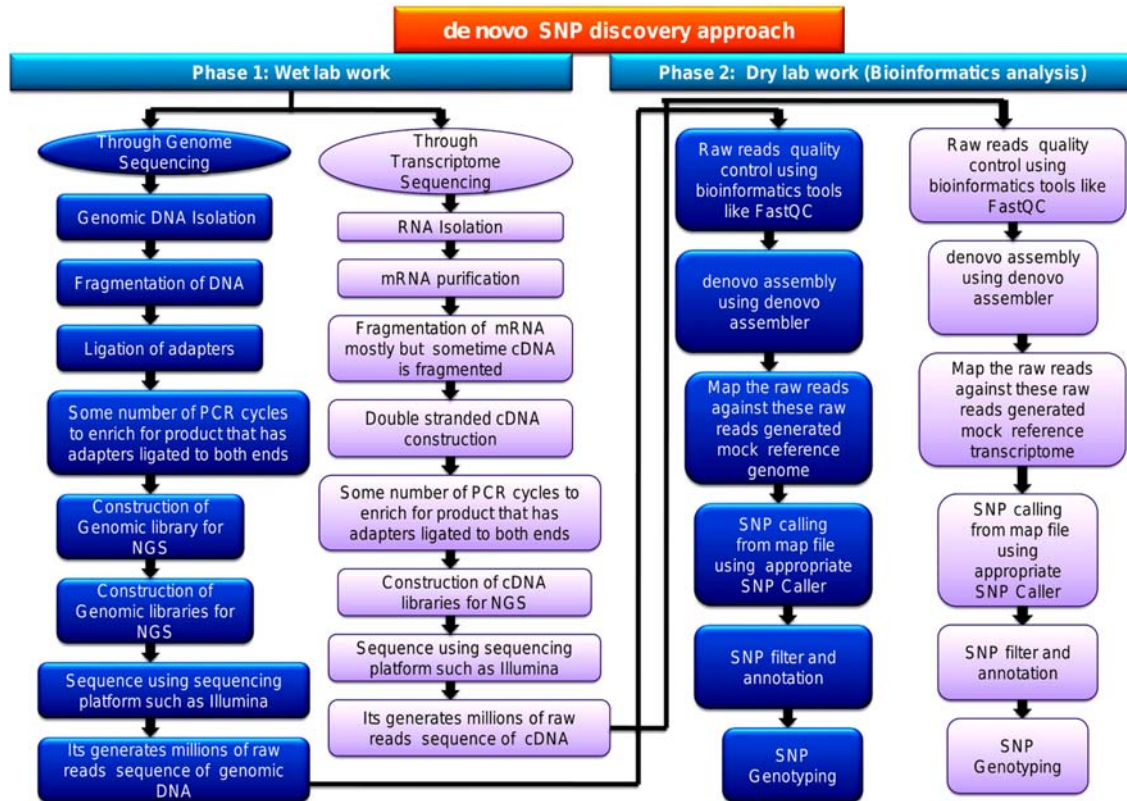


FIGURE 31.2 Workflow of SNP discovery and thereby its analysis through the de novo–based strategy. The generalized workflow from computational SNP discovery and can be applied for SNP discovery in plants that do not have deposited databases of reference genome/unigene/transcriptome. Thus it requires the creation of Mock reference from the same raw sequence reads. *SNP*, Single-nucleotide polymorphisms.

31.2.1.1 Sample preprocessing and DNA or RNA extraction

In the case of plants, it is necessary to handle the sample with proper care. The samples could be used for the DNA and/or RNA isolation using a specific protocol. After separation from the plants the peel tissues may be frozen in liquid nitrogen on the day of collection and stored at -80°C for further DNA or RNA extraction; otherwise, tissue can be used directly after peeling on the day of separation. The tissue could be ground to a fine powder in liquid nitrogen and the genomic DNA is isolated using an improved tab method (Doyle & Doyle, 1987) or by using any commercially available DNA extraction kit. Similarly, for RNA isolation, tissue could also be ground to a fine powder in liquid nitrogen and the total RNA could be isolated by using CTAB protocol (Chang, Puryear, & Cairney, 1993) or TRIzol Reagent (Rio, Ares, Hannon, & Nilsen, 2010) or by commercially available RNA extraction kit. The RNA sample concentrations can be estimated by using RNA high sensitivity Assay kit such as QubitRNA high sensitivity assay kit (Life Technologies, CA, the United States) and the RNA Nano 6000 assay kit can evaluate RNA integrity. DNA/RNA purity and quantity can be estimated with the NanodropBiospectrophotometer (Eppendorf, Germany).

31.2.1.2 Library preparation

The genomic libraries and cDNA libraries are constructed for DNA and RNA samples, respectively. RNA/DNA samples are extracted from fragmented sample tissue/cells. Through reverse transcription, RNA is converted to cDNA. DNA fragments are converted to the library by ligation with sequencing adapters containing specific sequences designed to interact with the NGS platform. The size of target DNA fragments in the final library is a key parameter for NGS library construction. Three approaches are available to fragment nucleic acid chains: physical, enzymatic, and chemical. DNA fragmentation is typically done by physical methods (acoustic shearing and sonication) or enzymatic methods (nonspecific endonuclease). The applications also determine the size of an RNA-Seq library.

Scientists typically do fundamental gene expression analysis using single-end 100 base reads. In most instances the RNA is fragmented before conversion into cDNA, which is typically done using controlled heated digestion of the

RNA with RNase in the presence of a divalent metal cation (magnesium or zinc) (Forconi & Herschlag, 2009). Once the starting DNA has been fragmented, the fragment ends are blunted and 5' ends are phosphorylated using a mixture of three enzymes: T4 polynucleotide kinase, T4 DNA polymerase, and Klenow large fragment. Next, the 3' ends are A-tailed using either *Taq* polymerase or Klenow fragment. During the adapter ligation the optimal adapter: fragment ratio is 10:1, calculated based on copy number or molarity. Too much adapter favors the formation of adapter dimers that can be difficult to separate and dominate in the subsequent PCR amplification.

For mRNA sequencing libraries, methods have been developed based on cDNA synthesis using random primers, oligo-dT primers, or by attaching adapters to mRNA fragments followed by some form of amplification. The mRNA can be primed by random oligomers or by an anchored oligo-dT to generate first-strand cDNA and it is converted into ds cDNA via PCR. The controlled heated digestion of RNA works well for 6–12 samples because this is a rapid process (5–10 min) but prone to overfragmentation if not well controlled (Forconi & Herschlag, 2009). Therefore it is difficult to use in an HTP platform. The NEBNext DNA fragmentase enzyme mix (New England Biolabs, United Kingdom) is available that can cleave ds cDNA molecules to overcome these problems, which provides more uniform libraries when processing numerous samples. Both molecular fragments are ligated into a suitable sequencing vector molecule for the next-generation platform (Head et al., 2014).

31.2.1.3 Next-generation sequencing

The constructed libraries are clonally amplified for sequencing. The sequencer utilizes these libraries separately of DNA and cDNA for genome and transcriptome sequencing, respectively (Berghlund, Kiialainen, & Syvänen, 2011). This machine generates a considerable amount of nucleic acid sequence called raw sequence data. Illumina, Roche/454 Life Sciences, Life Technologies/Applied Biosystems SOLiD, Life Technologies/Ion Torrent, and Helicos BioSciences are commonly used platforms next-generation platforms (Petersen & Coleman, 2020).

31.2.1.4 Quality control and alignment to the reference genome

All raw data obtained are further processed through filtering parameters. After removing reads containing adapter or ploy-N and low-quality reads from raw data, high-quality clean data collected in FASTQ format are used for reference-based assembly. In reference-based assembly, reads are assembled and align simultaneously. The alignment generates the map file, which is used for SNP calling. Reference-based assembly requires that the organism must have a suitable reference genome. The assembly tools are Trinity (Grabherr et al., 2011) IDBA-UD (Peng et al., 2013), Oases (Schulz, Zerbino, Vingron, & Birney, 2012), Trans-abyss (Simpson et al., 2009), etc.

31.2.1.5 Single-nucleotide polymorphism calling

The SNP calling aims to determine where positions there are polymorphisms or where positions at least one of the bases differ from a reference sequence; the latter is also sometimes referred to as variant calling or SNP. The accuracy of the alignment has a crucial role in variant detection. Incorrectly aligned reads may lead to errors in SNP and genotype calling, so alignment algorithms need to be able to cope with sequencing errors and potentially real differences between the reference genome and the sequenced genome that are due to polymorphisms. SNP caller such as ComB (Souaiaia, Frazier, & Chen, 2011), HaploSNPer (Tang, Leunissen, Voorrips, van der Linden, & Vosman, 2008), QualitySNP (Tang, Vosman, Voorrips, van der Linden, & Leunissen, 2006), SNP-PHAGE (Matukumalli et al., 2006), GTAK (McKenna et al., 2010), SNIPlay (Dereeper et al., 2011), SNIPlay3 (Dereeper et al., 2015), GBS-SNP-CROP (Melo, Bartaula, & Hale, 2016), UGbs-Flex (Qi et al., 2018), GB-eaSy (Wickland, Battu, Hudson, Diers, & Hudson, 2017), Fast-GBS (Torkamaneh, Laroche, Bastien, Abed, & Belzile, 2017), freebayes (Bian et al., 2018) SOAPsnp (Li et al., 2009), SAMtools (Li, 2011) can be utilized for the mining of SNPs from the map file.

31.2.2 De novo single-nucleotide polymorphism discovery

In de novo sequence data the grouping of the sequence data that belong to the same region of the genome, special assembly tools are employed to split up the input datasets that are not assembled as contigs. When the number of reads becomes too large, then the process consumes time. Specialized tools like d2cluster (Burke, Davison, & Hide, 1999), Teraclu and TGICL (Perteau et al., 2003) have been developed to perform initial segregation of sequence fragments into homologous groups, which are further decomposed into clusters of unique origin. After the clustering step, each cluster needs to be processed to align all reads within the cluster. All nucleotides from different reads simultaneously on the gene or genome are aligned and can be easily compared. If some fragments cannot be properly aligned, they do not

belong to a single cluster and are split into a second cluster. After individual reads have been clustered into aligned homologous groups, the final step of polymorphism identification is finding variations in the alignment file and applying a scoring scheme.

The generalized workflow for de novo–based SNP discovery is shown in Fig. 31.2. These two phases of de novo SNP discovery are described in detail in the following sections.

Sample preprocessing and DNA or RNA extraction, library preparation, and sequencing steps are the same as described earlier for the reference-based approach (Sections 31.2.1.1–31.2.1.3 and 31.2.1.5).

31.2.2.1 Quality control and de novo assembly

The quality control of raw reads is done as described in Section 31.2.1.4. In a de novo genome or transcriptome assembly and annotation project, the raw nucleotide sequence is assembled as completely as possible and then annotated with a nonredundant database (NR database). This assembled genome or transcriptome can be used as a reference for mapping or alignment. The assembly of raw sequence reads into contigs can be performed using assembler tools such as Velvet (Zerbino, 2010) SPAdes (Bankevich et al., 2012), Trinity (Grabherr et al., 2011), IDBA-Tran (Peng et al., 2013), Oases (Schulz et al., 2012), SOAPdenovo (Xie et al., 2014), trans-abyss (Simpson et al., 2009), Multiple-k (Surget-Groba & Montoya-Burgos, 2010; Martin et al., 2010), BinPacker (Liu, Yu, Jiang, & Li, 2016; Liu, Wu, Li, & Boerwinkle, 2016; Liu, Li, et al., 2016), TransLiG (Liu, Yu, Mu, & Li, 2019), etc.

31.2.2.2 Alignment or mapping of high-quality raw read to the mock reference genome

The high-quality raw read is mapped against the newly generated reference of genome or transcriptome. Read mapping aligns the reads on reference genomes. The alignment tools such as TopHat (Trapnell et al., 2009), Read-Split-Run (Bai, et al., 2016), and bowtie (Langmead, 2010) can take input of reference genome and a set of high-quality raw reads. These tools can align each read set on the reference genome, reading the mismatches and indels of some short fragments on the two ends of the reads. The mapping tools such as TopHat (Trapnell, Pachter, & Salzberg, 2009), Bowtie (Langmead, 2010), WIT (Kumar, Agarwal, & Ranvijay, 2019), SRmapper (Gontarz, Berger, & Wong, 2013), SRPRISM (Morgulis & Agarwala, 2020), HISEA (Khiste & Ilie, 2017), HISAT2 (Kim, Paggi, Park, Bennett, & Salzberg, 2019), BWA (Li & Durbin, 2009), Bowtie 2 (Langmead & Salzberg, 2012), BarraCUDA (Klus & Lam, 2012), RazerS 3 (Weese, Holtgrewe, & Reinert, 2012) are used for the generation of the map file. The generated map file is being utilized for the calling of SNPs by using SNP caller tools. Some commonly used SNP discovery tools such as sequence alignment, reference-based sequence assembly, de novo assembly, and SNP calling are listed in Table 31.1.

31.3 Single-nucleotide polymorphism annotation

It is a crucial step to transform newly discovered SNPs into meaningful information. These meaningful annotations of SNPs provide pivotal information such as the affected genes, the variant's effects at the protein products level, and the minor allele frequency. The output from the variant discovery tools or pipelines is a huge variant calling format (VCF) file containing hundreds of thousands of rows. The VCF file contains both SNPs and small insertions and deletions (INDELs). These VCF files are being annotated using several variant identification tools, such as ANNOVAR (Wang et al., 2010), SnpEff (Cingolani et al., 2012), AnnTools (Makarov et al., 2012), and PolyPhen (Thusberg, Olatubosun, & Vihinen, 2011). A dbNSFP tool is a combination of many programs such as SIFT, Polyphen2, LRT, PhyloP, and MutationTaster into a single database for annotation and filtering purposes (Liu, Jian, & Boerwinkle, 2011). Recently, dbNSFP has upgraded to include more databases and tools, that is, FATHMM, Mutation Assessor, and others, and also provides the ability to annotate splice-site variants as well (Liu, Li, et al., 2016; Liu, Wu, et al., 2016; Liu, Yu, et al., 2016). Many developed tools can annotate lists of genes regarding their biological product role and filter the considerable variant lists. However, such tools require advanced computational skills for both the installation process and their usage (Hart et al., 2016). There are web services available to annotate the SNPs. For instance, GoGene is a web service that can annotate the gene for an SNP using the gene ontology (GO) and Medical Subject Headings (MeSH) vocabularies (Plake, Royer, Winnenburger, Hakenberg, & Schroeder, 2009). Var2GO is the first web tool that permits to upload a complete raw variants file, annotate both the variants and the related genes, and interactively filter them, obtaining a reduced file with all the needed information (Granata, Sangiovanni, Maiorano, Miele, & Guarracino, 2016). There are also some command-line tools such as AnnoKey (Park, Nguyen-Dumont, Kang, Verspoor, & Pope, 2014) SNP2GO (Szkiba, Kapun, von Haeseler, & Gallach, 2014), which have been used to annotate the SNP variants. There are some more command-line tools like GEMINI (Paila, Chapman, Kirchner, & Quinlan, 2013), KGGSEQ (Li et al., 2012),

TABLE 31.1 List of single-nucleotide polymorphism (SNP) discovery tools, namely, sequence alignment, reference-based sequence assembly, de novo assembly, and SNP calling.

Process	Tool	References
Sequence alignment	TopHat 2	Kim et al. (2016)
	GSAAlign	Lin and Hsu (2020)
	AVID	Bray, Dubchak, and Pachter (2003)
	BLAST	Altschul, Gish, Miller, Myers, and Lipman (1990)
	BBBWT	Lippert (2005)
	BALT	Kent et al. (2002)
	BLASTZ	Schwartz et al. (2003)
	Cgaln	Nakato and Gotoh (2010)
	Chain Cleaner	Suarez et al. (2017)
	Harvest	Treangen et al. (2014)
	LAGAN	Brudno et al. (2003)
	LAST	Kielbasa, Wan, Sato, Horton, and Frith (2011)
	MAGIC	Swidan, Rocha, Shmoish, and Pinter (2006)
	MUMmer4	Marçais et al. (2018)
	Minimap2	Li (2018)
	MAQ	Li, Ruan, et al. (2008), Li, Li, et al. (2008)
	HISTAT	Kim, Langmead, and Salzberg (2015)
	HPG Aligner	Tárraga et al. (2014)
	Segemehl	Hoffmann et al. (2009)
	SSAHA	Ning et al. (2001)
	Kart	Lin and Hsu (2017)
	STAR	Dobin et al. (2013)
	SpliceMap	Au, Jiang, Lin, Xing, and Wong (2010)
MapSplice	Wang, Li, and Hakonarson (2010)	
GSNAP	Wu and Nacu (2010)	
Reference-based sequence assembly	Scallop	Shao and Kingsford (2017)
	TransComb	Liu, Li, et al. (2016), Liu, Yu, et al. (2016), Liu, Wu, et al. (2016)
	StringTie	Pertea et al. (2015)
	RaGoo	Alonge et al. (2019)
	RGAAT	Liu et al. (2018)
	RECORD	Buza, Wilczynski, and Dojer (2015)
	Cufflinks	Trapnell et al. (2010)
	Bayesember	Marett, Sibbesen, and Krogh (2014)
	Isolnfer	Feng, Li, and Jiang (2011)
	IsoLasso	Li, Feng, and Jiang (2011)
	iReckon	Mezlini et al. (2013)
	CEM	Li, Gui, Kwan, Bao, and Sham (2012)
CIDANE	Canzar, Andreotti, Weese, Reinert, and Klau (2016)	

(Continued)

TABLE 31.1 (Continued)

Process	Tool	References
De novo assembly	BinPacker	Liu, Li, et al. (2016), Liu, Yu, et al. (2016), Liu, Wu, et al. (2016)
	Bridger	Chang et al. (2015)
	Trinity	Grabherr et al. (2011)
	IDBA-Tran	Peng et al. (2013)
	SOAPdenovo-Trans	Xie et al. (2014)
	ABYSS	Simpson et al. (2009)
	Oases	Schulz et al. (2012)
	Velvet	Zerbino (2010)
SNP calling	VarSome	Kopanos et al. (2019)
	GATK	McKenna et al. (2010)
	BreakDancer	Chen et al. (2009)
	SOAP	Li et al. (2008)
	VarScan	Koboldt et al. (2009)
	Samtools	Li (2011)
	SnEff	Cingolani et al. (2012)
	CoVaCS	Chiara et al. (2018)
	AMLVaran	Wünsch, Banck, Müller-Tidow, and Dugas (2020)
	Platypus	Rimmer et al. (2014)
	SNVer	Wei, Wang, Hu, Lyon, and Hakonarson (2011)
	VarDict	Lai et al. (2016)
	LoFreq	Wilm et al. (2012)
	QCALL	Le and Durbin (2011)
	Dindel	Albers et al. (2011)
	Atlas-SNP2	Challis et al. (2012)
	CRISP	Bansal (2010)
	SeqEM	Martin et al. (2010)
	SLIDERII	Malhis and Jones (2010)
	SNP-o-matic	Manske and Kwiatkowski (2009)
SOAP2	Li et al. (2009)	
mrsFASTultra	Hach et al. (2014)	

Plink/SEQ (Plink/SEQ Home Page), which are more focused on variant rather than gene annotation, and all exclusively usable from the command-line interface.

31.4 Single-nucleotide polymorphism database

The rapid development in HTP genotyping has opened up new avenues in genetics while at the same time producing immense data handling issues and efficient data storage and manipulation of SNP genotypes, and access by multiple users are the major issues. Therefore many of the databases have been developed to address some of the issues

TABLE 31.2 Some common single-nucleotide polymorphism (SNP) databases specific to plants.

SNP database	Organism	References
dbSNP	Commonly for all organisms but this has not been significantly adopted by researchers who work on nonhuman species	https://www.ncbi.nlm.nih.gov/snp/
Panzea	Maize	Zhao et al. (2006)
Triticeae toolbox	Small grain crops	Blake et al. (2016)
GrainGenes	Small grain crops	Matthews, Carollo, Lazo, and Anderson (2003)
SorGSD	Sorghum	Luo et al. (2016)
CropSNPdb	Wheat and brassica SNPs array data	Scheben et al. (2019)
POLYMORPH website	<i>Arabidopsis thaliana</i>	http://polymorph.weigelworld.org/cgi-bin/retrieve_snp.cgi
SNiPlay	Currently database of 4 Vitis projects and one <i>Coffea</i> project	Dereeper et al. (2011)
AutoSNPdb	Currently containing database barley, brassica, rice, and wheat	Duran et al. (2009)
The maritime pine SNP database	Contains information of SNPs in the ESTs in pine trees	Le Dantec et al. (2004)
ESTree DB	SNP report of peach tree and almond	Lazzari et al. (2005, 2008)
PlantMarkers database	For both plants and animals	Rudd, Schoof, and Mayer (2005)
TreeSNPs	For plants	Clément, Fillon, Bousquet, and Beaulieu (2010)
The BGI-RIS database	For rice	Zhao et al. (2004)
IRIS	For rice	Bruskiewich et al. (2003)
Orygenes DB	For rice	Droc, Perin, Fromentin, and Larmande (2009)
EUSNPDB	Eucalyptus	
QualitySNP	For plants	Tang et al. (2006)
PoMaMo	For potato	http://www.gabipd.org/projects/Pomamo
TAED	Comparative genomics data of plants	http://www.bioinfo.no/tools/TAED
Plant genome central	For plant	http://www.ncbi.nlm.nih.gov/genomes/plants/plantlist.html

(Fong et al., 2010; Mitha et al., 2011; Orro, Guffanti, Salvi, Macciardi, & Milanesi, 2008) while focusing on major SNP databases available for plants. In recent times a wide range of online and freely accessible databases are devised to recognize SNPs in genomic sequences. The lack of genomic databases designed to host crop SNP array data may contribute to the poor data availability after publication. Hence, the National Center for Biotechnology Information (NCBI) hosts the generic database dbSNP. However, researchers who work on nonhuman species have not significantly adopted this, and recently, NCBI decided to phase out support for nonhuman organisms. Crop research communities maintain SNP data in specialized databases, such as Panzea for maize (Zhao et al., 2006) and Triticeae toolbox for small grain crops (Blake et al., 2016). Some common SNP databases specific to plants are listed in Table 31.2.

31.5 Single-nucleotide polymorphism genotyping

There are numbers of methods available for genotyping of SNP but all of them are not equally useful. SNP genotyping relies on the ability to distinguish a single-base match from a single-base mismatch. There are many methods such as

minisequencing, electrophoresis and fluorescence detection, molecular beacons, array hybridization, fluorescence detection, MALDI-TOF MS, fluorescence polarization, pyrosequencing, microarrays, and fluorescence detection and invasive cleavage, allele-specific PCR. Technological advancements have made SNP genotyping efficient as well as automated.

The SNP genotyping can be broadly classified into two groups, gel- and nongel-based assays.

31.5.1 Gel-based single-nucleotide polymorphism genotyping

There are three gel-based methods available for SNP genotyping.

31.5.1.1 Cleaved amplified polymorphic sequence markers

The first one is the RFLP assay, in which SNP can be detected by RFLP of PCR products whenever the presence of SNP eliminates the restriction site for a particular restriction enzyme. After the digestion of the PCR product, it is subjected to RFLP to detect the differences in patterns that will be due to SNP and such markers are called cleaved amplified polymorphic sequence (CAPS) markers (Fig. 31.3).

31.5.1.2 Single-stranded conformation polymorphism

The second method of gel-based SNP genotyping is single-stranded conformation polymorphism (SSCP) is based on the DNA conformation (Fig. 31.4).

The third method of gel-based SNP genotyping is allele-specific amplification. It is based on distinguishing between two DNA targets differing at one nucleotide position by hybridization (Wallace et al., 1979). Two allele-specific probes are designed with the polymorphic base in the central position of the probe sequence. Only the perfectly matched probe-target hybrids remain stable under optimized assay conditions and hybrids with one-base mismatch are unstable. Allele-specific probes (AS probe) with reverse dot-blot formats were used to detect the first polymorphisms analyzed by PCR in the agriculture field and they are still used in some laboratories (Fig. 31.5). To take full advantage of new AS probe formats for SNP typing, it is necessary to use detection methods that provide high accuracy, high sensitivity, and HTP (Kim et al., 2016).

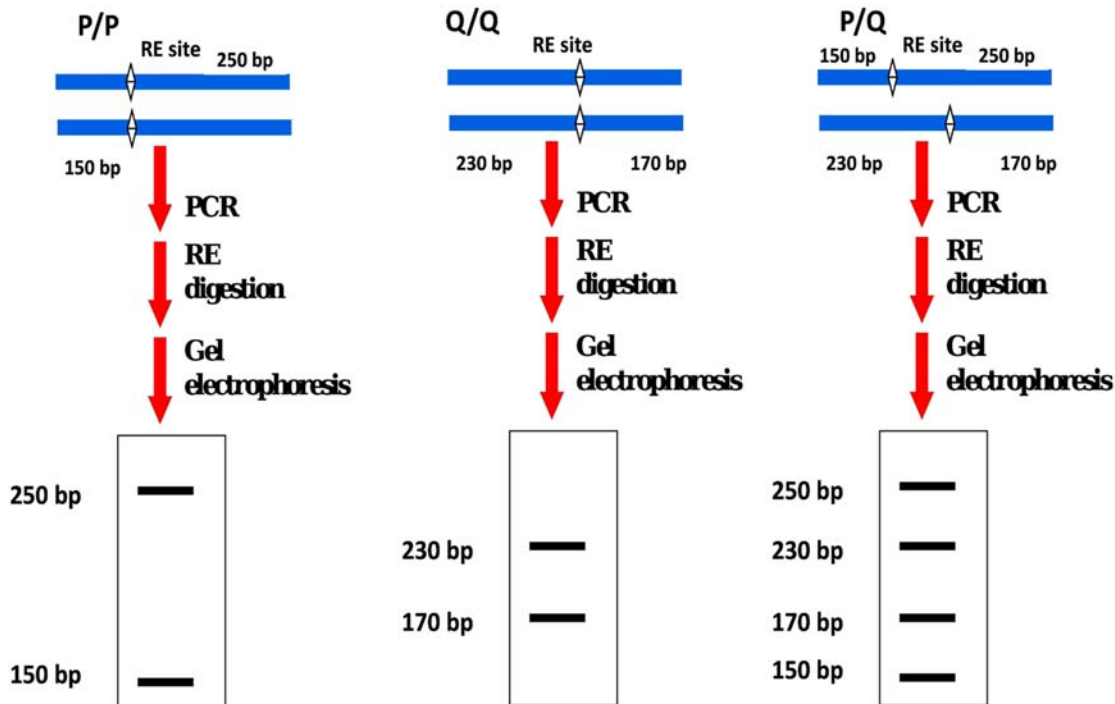


FIGURE 31.3 Representation of cleaved amplified polymorphic sequence marker system for single-nucleotide polymorphisms genotyping.

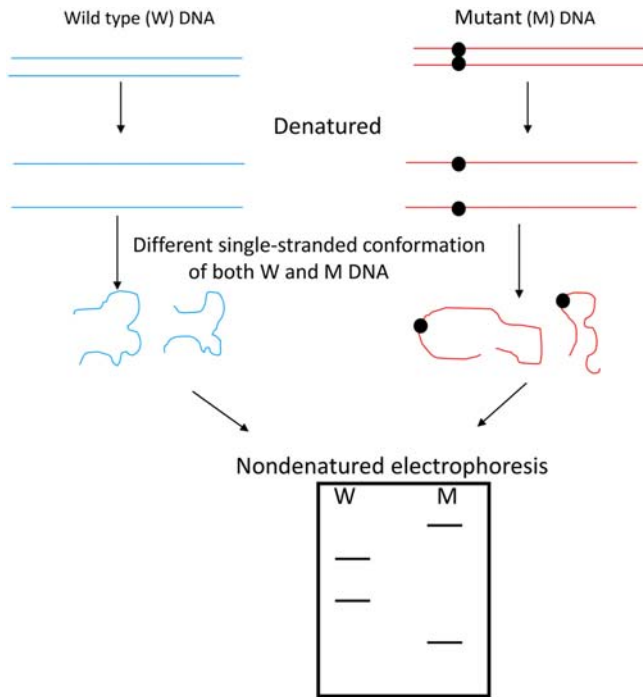


FIGURE 31.4 The polymerase-chain reaction–based single-stranded conformation polymorphism analysis. The single-nucleotide polymorphisms (represented by a dot on a DNA strand) lead to different conformations of single-strand for the mutant DNA (M) compared with the wild type (w) and thus resulting in differential mobilities in nondenaturing gel electrophoresis.

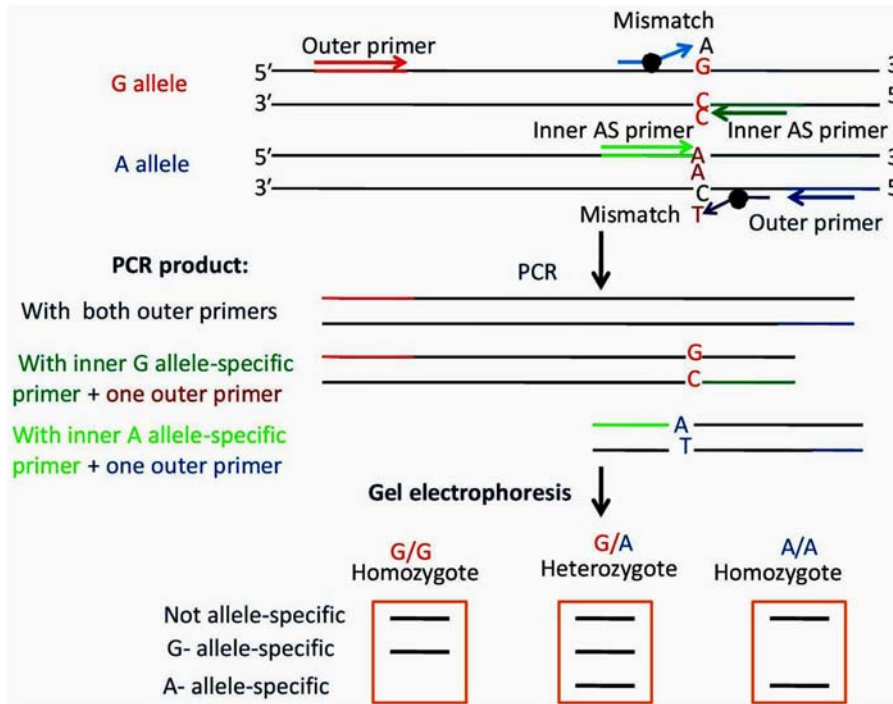


FIGURE 31.5 Schematic illustration of the tetra-primer allele-specific polymerase-chain reaction for single-nucleotide polymorphism genotyping. Upper two allele-specific amplicons are generated using two pairs of primers, one pair (outer forward and reverse inner) producing an amplicon representing the nonallele-specific product. The specificity of the inner primers is conferred by two mismatches, one between the 3' terminal base of an inner primer and the template and the second at position-2 from the 3' terminus (indicated by an asterisk). Lower by positioning the two outer primers at different distances from the polymorphic nucleotide, the two allele-specific amplicons differ in length, allowing them to be discriminated by gel electrophoresis.

31.5.2 Nongel-based single-nucleotide polymorphism genotyping

The nongel-based method for SNP genotyping requires PCR amplification; however, the amplified product is detected through the nongel techniques that will discriminate between the wild and mutant alleles. The nongel techniques detect a mismatch or a perfect match between the amplified product and oligonucleotide probes. Several nongel-based methods like minisequencing, fluorescence detection, molecular beacons, MALDI-TOF MS (matrix-assisted laser

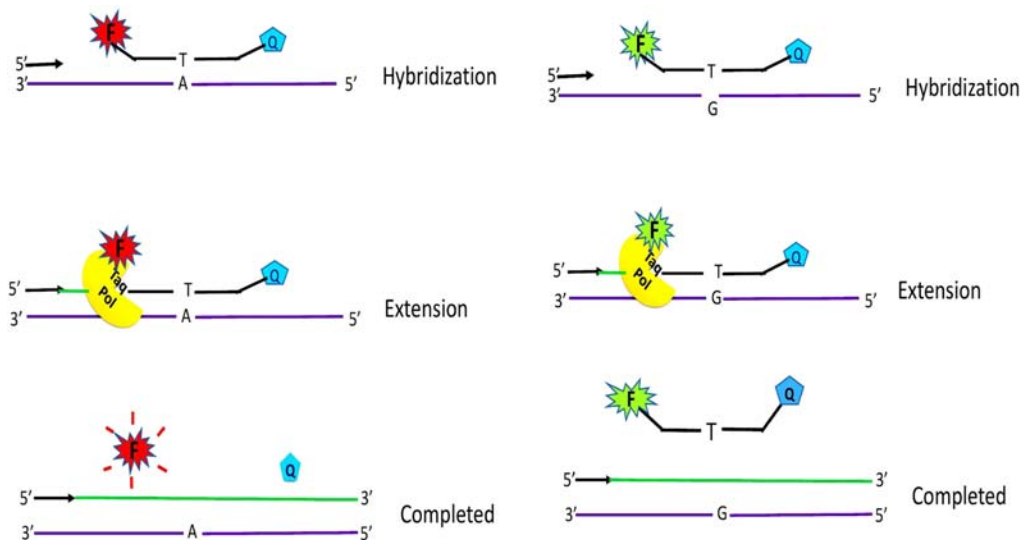


FIGURE 31.6 Probe binding and primer extension in a TaqMan SNP Genotyping assay. Allelic discrimination is achieved by the selective annealing of matching probe and template sequences, which generates an allele-specific (fluorescent dye-specific) signal.

desorption/ionization time-of-flight mass spectrometry), fluorescence polarization, pyrosequencing, microarrays, and invasive cleavage (Gupta, Roy, & Prasad, 2001). Two commonly used techniques, TaqMan assay and minisequencing for SNP genotyping, are being discussed.

31.5.2.1 TaqMan assay

The TaqMan assay is based on the 5' nuclease activity of Taq polymerase that displaces and cleaves the oligonucleotide probes hybridized to the target DNA, generating a fluorescent signal (Livak, Flood, Marmaro, Giusti, & Deetz, 1995). Two TaqMan probes differing at the polymorphic site are needed: one probe is complementary to the wild-type allele and the other to the variant allele. These probes have different fluorescent dyes attached at the 5' end and a quencher attached at the 3' end (Livak, 1999). If there is a match between the primer–probe and target DNA, then it will be wild type and if there is a mismatch, then it will be the mutant allele having the SNP (Fig. 31.6).

31.5.2.2 Minisequencing

SNPs can be validated by sequencing the few bases around the SNP site through Sanger's dideoxynucleotide methods. In this reaction a primer that anneals to its target DNA immediately adjacent to the SNP is extended by a DNA polymerase with a single-nucleotide complementary to the polymorphic site (Syvänen, Aalto-Setälä, Harju, Kontula, & Söderlund, 1990). Different technologies are available for analyzing primer extension products. The use of labeled or unlabeled nucleotides, ddNTP combined with dNTP, or only ddNTP in the minisequencing reaction depends on the product detection method. The multiplexing ability also relies on the technology used. The most common technologies used for analyzing minisequencing products are fluorescence detection, MALDI-TOF MS, and microarrays.

31.6 Application of single-nucleotide polymorphisms in plants

Major impediments in meeting global food demands in light of the growing world population include deterioration and shrinkage of agricultural lands due to salinization, environmental pollution, temperature, global climate change, urban and industrial growth, etc. (FAO, 2020). Therefore the development of new genotypes that could grow in deteriorated/marginal/waste agricultural lands is the need of hour. In this context the novel crop improvement approaches are highly desirable (Lateef, 2015). Thus identification of SNPs which enables the selection of desired lines in large-scale populations is of immense importance in crop breeding programs in an efficient and economical way (Brennan et al., 2014).

As the genome of many plants has been fully sequenced, the discovery of interest-specific sequence differences becomes easier (Fournier-Level et al., 2013). The application of SNPs in detecting relationships between allelic forms of a gene and phenotypes, prevalent in diseases with multifactor genetics, high-resolution genetic map construction, linkage disequilibrium–based association mapping, genetic diagnostics, genetic diversity analysis, cultivar

identification, phylogenetic analysis, etc., creates great potential for characterization of genetic resources (Freudenthal, Ankenbrand, Grimm, & Korte, 2019). Thus in the following sections, some common applications of the SNP marker in the field of agriculture are discussed in detail.

31.6.1 Genetic diversity

Information on genetic diversity and relationships among lines and varieties is important to plant breeders for the improvement of crop plants. Knowledge of genetic diversity is also valuable for identifying novel alleles, which may then be introgressed into elite backgrounds within breeding programs. Previously, assessing diversity on a genome-wide scale was based on marker systems such as AFLPs, SSRs, or isozymes (Vigouroux et al., 2005). However, with HTP SNP technology development, it has become possible to assess diversity within specific genes or genomic regions more efficiently and reliably. Thus in maize, genetic diversity was studied using SNPs at 21 loci along chromosome 1 (Tenaillon et al., 2002). This study facilitated an understanding of the forces contributing to genetic diversity in maize.

Similarly, genetic polymorphism studies have been performed among Tongkat Ali's different accessions (*Eurycoma longifolia*), a plant used in herbal remedies and health supplements through allele-specific oligonucleotide hybridization (Osman et al., 2003). On average 64% loci were found polymorphic, and the populations were found to exhibit a high degree of diversity. SNPs have also been used for cultivar identification in malting barley and wheat cultivars (Batley & Edwards, 2007). These assays could also be applied for distinctness, uniformity, and stability testing and assessing plant breeder's rights (Chiapparino, Lee, & Donini, 2004). Van Inghelandt, Melchinger, Lebreton, and Stich (2010) studied the comparative diversity analysis based on SNP markers among the maize varieties and obtained 4 (Iodent and SSS) to 25 (Flint) group-specific alleles with a gene diversity index of 0.32. Singh, Choudhury, et al. (2013), Singh, Gupta, et al. (2013) investigated comparative genetic diversity analysis between SSR and SNP markers among the Indian rice varieties. It was observed that SNP markers detected a greater extent of variation (45.2% for SNP and 13.3% for SSR) and classified the rice accessions more accurately. Du et al. (2019) investigated the genetic diversity among varieties and subpopulations of pepper (*Capsicum spp.*) using SNP marker that classified the pepper populations into various groups, namely, long horn-fruited followed by the shorthorn-, linear-, and blocky-fruited populations. Xia et al. (2019) studied the population structure analysis based on the SNP markers of 200 individuals of African oil palm (*Elaeis guineensis*) and classified them into five subgroups. Adawiah, Norliza, Fairuz, Norzihan, and Kalsom (2016) discovered the 934 and 7959 putative SNPs in Eksotika and Sekaki, varieties of papaya, respectively, and these can be utilized for the genetic diversity analysis and variety identification.

31.6.2 Genetic mapping

The whole-genome trait mapping by allele association requires high marker density, which SNPs can readily provide. In combination with their HTP discovery and detection methodology, the abundance of SNPs makes them suitable markers of choice for applications such as linkage mapping, QTL mapping, and association mapping. SNP-based genetic mapping has been demonstrated both on large and small scale, in both models (Schmid et al., 2003) and nonmodel (rice, wheat, pepper, date palm, cannabis, potato, tomato, etc.) plants (Carpentier et al., 2008). SNPs identified within ESTs or large genomic fragments can be applied for genetic mapping of complex traits. This approach enables the genetic mapping of specific genes of interest and assists in identifying linked or perfect markers for traits and increasing the density of markers on genetic maps (Rafalski, 2002). SNP markers also allow the integration of genetic and physical maps. SNPs can also develop haplotyping systems for genes or regions of interest (Delourme et al., 2013). A genome-wide set of SNP markers in *Arabidopsis thaliana* has been identified (Schmid et al., 2003). Alternatively, a targeted approach may be undertaken to map candidate genes, or the fine mapping of specific genomic regions that may have previously been identified through QTL mapping and SNPs has also been used to remap the genetically mapped genes (Ching & Rafalski, 2002). The abundance of SNPs makes them useful for placing ESTs or candidate genes onto a genetic map, which has been previously constructed with other markers. Thus SNPs have been identified and characterized in soybean ESTs which have been used to develop soybean linkage maps for its association with quantitative trait loci (Zhu et al., 2003).

Identification of EST-based SNPs associated with traits of interest in barley has been made for the syntenic studies with other related species (Kota, Varshney, Thiel, Dehmer, & Graner, 2001). Similarly, SNPs have been genetically mapped in melon (Morales, Roig, Monforte, Arus, & Garcia-Mas, 2004) and cassava (Lopez et al., 2005). SNPs have also been applied in maize to generate a high-resolution genetic map (Sanchez-Villeda et al., 2003). Yu et al. (2011) analyzed a comparative study of SNP-based map to that of a previously generated RFLP/SSR-based QTL map in rice

populations for traits such as yield, number of tillers per plant, number of grains per panicle, and grain weight. SNP-based maps were found to be denser than the RFLP/SSR-based QTL maps. In pea, 64,754 SNPs have been analyzed by short-read sequencing, which was used to construct a whole-genome genotyping by sequencing (WGGBS)-derived pea genetic map, which was found to be collinear with previous pea consensus maps (Boutet et al., 2016). Geleta, Gustafsson, Glaubitz, and Ortiz (2020) constructed the genetic linkage map of *Lepidium campestre* based on 2330 SNP markers and the final linkage map consisted of eight linkage groups.

31.6.3 Phylogenetic analysis

Plant phylogenetic and evolutionary studies have traditionally relied on sequence diversity, and therefore SNPs are the most interesting than all other sequence variations due to their dense abundance in the genome. Nuclear and chloroplast genes are a rich source of phylogenetic information for evolutionary analysis in plants. The diversity of the sequence and genotyping of these SNPs can be used to infer phylogenetic and evolutionary relationships in a wide variety of species. Genetic inheritance studies could be deduced through analysis of SNP diversity and conservation among sequences from individuals. By considering mutation rates a molecular clock may also be applied to estimate the timing of species divergence. Increasing quantities of sequence and SNP data for genes in a wide variety of species is slowly uncovering the molecular mechanisms of evolution within genomes and between species. It is possible to utilize other molecular markers for phylogenetic analysis. However, without the knowledge of the sequence variation, degrees of similarity only can be assessed and homoplasy cannot be ruled out (Paule et al., 2020). Shavrukov et al. (2014) used the 863 SNP markers for the phylogenetic analysis in bread wheat from Kazakhstan and they identified as unique to specific cultivars, and clusters of these markers showed specific patterns on the consensus genetic map for each cultivar.

Intervarietal polymorphism-based phylogenetic analysis showed that the ancient cultivar *Erythrospermum* 841 was the most genetically distinct from Kazakhstan's other nine cultivars, falling in a clade with the American cultivar Sonora and Central and South Asian genotypes. The modern cultivar, Kazakhstanskaya 19, also belongs to a separate clade, together with the American cultivar, Thatcher. Remaining eight cultivars share a single subclade, categorized into four clusters. The phylogenetic relationships among 199 accessions of chrysanthemum (*Chrysanthemum morifolium* Ramat.) have been performed based on 92,830 SNPs. All the accessions were found to be grouped into five clades (Chong et al., 2016). Acquadro et al. (2017) analyzed the phylogenetic relationship between the 76 accessions of eggplant species (*Solanum sp.*). All the 76 accessions were found to be clustered into four clades.

31.6.4 Marker-assisted selection

Functional genomic approaches such as transcriptomics, targeting-induced local lesions in genomes (TILLING), homologous recombinant, association mapping, and allele mining are all strategies to identify functional markers for breeding goals, such as agronomic traits and biotic and abiotic stress resistance (Salgotra & Stewart, 2020). Molecular markers are essential for mapping candidate genes, marker-assisted breeding, and the map-based cloning of genes underlying traits. Marker-assisted selection (MAS) is the often used or more prominently used genetic marker in plant breeding programs, which allows the breeder to achieve an early selection of a trait or a combination of traits. Molecular markers are 100% inheritable to the progeny, therefore using these markers to select for allow heritable trait is more effective and less expensive than phenotypic selection for that trait. The abundance of SNPs in plant genomes makes them attractive tools for MAS and map-based cloning and SNPs and indel molecular markers can be applied for MAS (Carrillo-Perdomo et al., 2020).

Markers loosely linked to a trait may suffer from recombination between the marker and the gene. Linked markers are also not usually transferable between populations originating from different parents due to a lack of polymorphism. Markers within the gene responsible for the trait are considered perfect markers. These are highly valuable for breeding as recombination between the marker and gene is practically eliminated, which is frequently transferable between populations.

SNPs are highly stable markers that may contribute directly to phenotype and they can serve as a powerful tool for MAS. Once SNP markers are found to be associated with a target trait, they can be applied by plant breeders for MAS to identify individual plants containing a combination of alleles of interest from large segregating populations. SNPs can be identified within or close to genes underlying agronomic traits. Although the SNP may not be responsible for the mutant phenotype, they may be applied for MAS and the positional gene cloning in the desired region of genome (Gupta et al., 2001). Association of SNPs with genes of economic value has already been demonstrated, such as SNP markers for supernodulation in soybean have been identified and the identified SNP in the GmNARK gene has been

suggested as a marker for hypernodulating mutation. The SNP was converted to a single-nucleotide amplified polymorphism marker to allow direct MAS for supernodulation at an early growth stage without inoculating and phenotyping the roots (Kim, Van, Lestari, Moon, & Lee, 2005). EST-based SNPs are associated with the *Adh* genes encoding alcohol dehydrogenase involved in the glycolytic pathway, which provides an ideal model for SNP discovery and analysis that can be used for genetic mapping and QTL analysis and MAS in sugarcane (Grivet, Glaszmann, Vincentz, Da Silva, & Arruda, 2003). An HTP SNP genotyping system has been developed and used to select barley alleles carrying superior alleles of β -amylase, a key enzyme involved in the degradation of starch during the malting process. The four allelic forms of the enzyme were unambiguously identified by genotyping two SNPs using the SnuPE system. A CAPS marker has also been developed, enabling the marker transfer to other laboratories that do not have SnuPE assay capabilities. These assays provide a rapid and inexpensive method for screening large numbers of individual plants, allowing the desirable allelic introgression into breeding programs (Paris, Jones, & Eglinton, 2002).

Further works on MAS using SNPs in barley include identifying SNPs in the *Isa* gene, which has a potential role in defense against pathogens. This gene was sequenced and screened for SNPs across 16 genotypes. This study showed little diversity in cultivated barley and that SNPs could be a useful tool for the introduction of novel alleles from wild barley (Bundock & Henry, 2004). Furthermore, SNPs associated with grain germination have been characterized across 23 varieties for their suitability for MAS implementation (Russell et al., 2004).

An SNP marker has been developed for the waxy gene controlling amylose content in rice. Amylose is the main component controlling the cooking and nutritional properties of cereals. Low amylose varieties are considered desirable, and in rice, it has been shown that the high and low amylose types can be differentiated based on an SNP near the waxy gene. This marker can be applied for MAS for the low amylose trait in seedlings (Gupta et al., 2001). Additionally, SNPs associated with essential genes in rice include an SNP marker for the dwarfing gene. The SNP was identified within an SSR flanking sequence and used for selection in various crosses. SNP-based markers for rice-blast resistance genes have also been developed. These markers enabled mapping the *Piz* and *Piz-t* genes, demonstrating that the SNPs are a valuable tool for gene mapping, map-based cloning, and MAS in rice (Hayashi, Hashimoto, Daigen, & Ashikawa, 2004).

In wheat the SNP linked to the protein structure of adenine phosphoribosyltransferase has been identified. This gene encodes the key enzyme, which converts adenine to adenosine monophosphate in the purine salvage pathway (Xing et al., 2005). In wheat, further SNPs in genes of interest have been identified, including the *Lr1* leaf rust resistance gene. Infections can lead to severe yield losses and therefore the desire is to grow resistant cultivars. The SNP marker development in the *Lr1* gene has exhibited a dramatic improvement on the STS markers, which was not previously specific in 50% of cultivars tested. The growing number of wheat SNP markers available can open the possibility of introducing multiplexed assays, targeting loci to pyramid trait selection during wheat breeding (Tyrka et al., 2004). Beukert et al. (2020) performed the MAS for improving rust resistance in hybrid wheat and thus findings suggested that MAS seemed to be a robust and efficient tool to improve leaf rust resistance in European wheat hybrids. Genome-wide association studies (GWAS) identified SNP marker–trait associations (MTAs) for 10 traits across the genome of Foxtail millet (*Setaria italica*). High-confidence MTAs for three crucial agronomic traits, including FLW (flag leaf width), GY (grain yield), and TGW (thousand-grain weight) were identified. The significant pyramiding effect of identified MTAs further supplemented its importance in breeding programs. Desirable alleles and superior genotypes were identified for foxtail millet improvement through MAS.

Work has also been performed on MAS in less developed crop species. A number of 132 SNPs in quinoa have been identified from ESTs. It was found that the SNP development from ESTs was a practical method for developing species-specific markers and may provide the molecular differentiation required to monitor gene flow between cultivated quinoa and weedy species (Coles et al., 2005). Further potential applications in plants include a study of nucleotide diversity in the *pall* locus of Scots pine. This gene is predicted to be associated with ozone tolerance, pathogen defense, and metabolism of exogenous compounds, and SNPs within it could prove valuable for MAS in this species (Dvornyk, Sirviö, Mikkonen, & Savolainen, 2002).

Qi, Talukder, Hulke, and Foley (2017) analyzed the SNP marker–based segregation of two downy mildew disease resistance genes, *PIArg* and *PI8*, which are highly effective against the causal fungus, *Plasmopara halstedii* races in sunflower. The rust caused by the fungus *Puccinia helianthi* and downy mildew by obligate pathogen, *P. halstedii*, are two of the most globally essential sunflower diseases. Four rust-resistant lines, HA-R3 (carrying the R_4 gene), HA-R2 (R_5), HA-R8 (R_{15}), and RHA 397 (R_{13b}), were each crossed with a standard line, RHA 464, carrying a rust gene R_{12} and a downy mildew gene PI_{Arg} , an additional cross of HA-R8 and RHA 397. Codominant SSR and SNP markers linked to the target genes were used to discriminate between homozygotes and heterozygotes in F_2 populations (Qi & Ma, 2020). Trebbi et al. (2019) analyzed the association between SNP markers and seed toxicity in the *Jatropha curcas*

L populations. The association study identified two new SNPs, SNP_J22 and SNP_J24, significantly linked to low toxicity with R^2 values of 0.75 and 0.54. It was suggested that these two valuable SNP markers could be used for HTP, marker-assisted breeding of seed toxicity in *J. curcas*.

31.7 Conclusion and prospects

Analysis of genetic diversity occurring within a given population is of significance for crop improvement. A variety of molecular markers have been used for plant genetic diversity analysis. The SNPs are a significant component of crop genomic diversity and are invaluable tools as genetic markers in research, breeding programs, and locating the genes associated with plant-specific traits. With the availability of various computational approaches and advancements in NGS technologies, SNP discovery became faster, efficient, and cost-effective, resulting in the identification of a large number of genome-wide SNPs from many plant species. Furthermore, technological advancements have made SNP genotyping more efficient as well as automated. Hence, SNPs are increasingly becoming the marker of choice for a wide range of applications, including genetic mapping, SNP marker–assisted plant breeding and diversity analysis.

Acknowledgment

The financial assistance in the form of research project sanctioned by Council of Science and Technology (CST), U.P., Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India, New Delhi, India, is gratefully acknowledged. Financial support from the Department of Higher Education, Government of U.P. under Institute for Development of Advanced Computing, ONGC Centre for Advanced Studies; Center of Excellence in Bioinformatics program, U.P.; Department of Biotechnology (DBT), Government of India, New Delhi; DST-PURSE grant is also gratefully acknowledged. The financial assistance in the form of research fellowships of SRF (to Dileep Kumar) and JRF (to Ranjana Gautam) by the University Grant Commission (UGC), New Delhi is gratefully acknowledged.

References

- Acquadro, A., Barchi, L., Gramazio, P., Portis, E., Vilanova, S., Comino, C., . . . Lanteri, S. (2017). Coding SNPs analysis highlights genetic relationships and evolution pattern in eggplant complexes. *PLoS one*, *12*(7), e0180774. Available from <https://doi.org/10.1371/journal.pone.0180774>.
- Adawiah, Z. R., Norliza, A. B., Fairuz, Y. M., Norzihan, A., & Kalsom, A. U. (2016). Sequence information on single nucleotide polymorphism (SNP) through genome sequencing analysis of *Carica papaya* variety Eksotika and Sekaki. *Journal of Tropical Agriculture and Food Science*, *44*(2), 219–228.
- Ahmad, T., Valentovic, M. A., & Rankin, G. O. (2018). Effects of cytochrome P450 single nucleotide polymorphisms on methadone metabolism and pharmacodynamics. *Biochemical Pharmacology*, *153*, 196–204.
- Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2011). Dindel: Accurate indel calls from short-read data. *Genome Research*, *21*(6), 961–973.
- Allen, A. M., Barker, G. L., Berry, S. T., Coghill, J. A., Gwilliam, R., Kirby, S., & Edwards, K. J. (2011). Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, *9*(9), 1086–1099.
- Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., & Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1), 1–17.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.
- Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., & Lander, E. S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, *407*(6803), 513–516.
- Arif, I. A., Bakir, M. A., Khan, H. A., Al Farhan, A. H., Al Homaidan, A. A., Bahkali, A. H., & Shobrak, M. (2010). A brief review of molecular techniques to assess plant diversity. *International Journal of Molecular Sciences*, *11*(5), 2079–2096.
- Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, *38*(14), 4570–4578.
- Bai, Y., Kinne, J., Donham, B., Jiang, F., Ding, L., Hassler, J. R., & Kaufman, R. J. (2016). Read-Split-Run: An improved bioinformatics pipeline for identification of genome-wide non-canonical spliced regions using RNA-Seq data. *BMC Genomics*, *17*(7), 107–117.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477.
- Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics (Oxford, England)*, *26*(12), i318–i324.
- Barchi, L., Lanteri, S., Portis, E., Acquadro, A., Valè, G., Toppino, L., & Rotino, G. L. (2011). Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics*, *12*(1), 1–9.
- Batley, J., & Edwards, D. (2007). *SNP applications in plants. Association mapping in plants* (pp. 95–102). New York, NY: Springer.

- Berger, J., Suzuki, T., Senti, K. A., Stubbs, J., Schaffner, G., & Dickson, B. J. (2001). Genetic mapping with SNP markers in *Drosophila*. *Nature Genetics*, 29(4), 475–481.
- Berglund, E. C., Kiialainen, A., & Syvänen, A. C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, 2(1), 1–15.
- Beukert, U., Thorwarth, P., Zhao, Y., Longin, C. F. H., Serfling, A., Ordon, F., & Reif, J. C. (2020). Comparing the potential of marker-assisted selection and genomic prediction for improving rust resistance in hybrid wheat. *Frontiers in Plant Science*, 11.
- Bian, X., Zhu, B., Wang, M., Hu, Y., Chen, Q., Nguyen, C., . . . Meerzaman, D. (2018). Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics*, 19(1), 1–11.
- Blake, V. C., Birkett, C., Matthews, D. E., Hane, D. L., Bradbury, P., & Jannink, J. L. (2016). The triticeae toolbox: Combining phenotype and genotype data to advance small-grains breeding. *The Plant Genome*, 9(2), plantgenome2014-12.
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3), 314.
- Boutet, G., Carvalho, S. A., Falque, M., Peterlongo, P., Lhuillier, E., Bouchez, O., & Baranger, A. (2016). SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. *BMC Genomics*, 17(1), 1–14.
- Bray, N., Dubchak, I., & Pachter, L. (2003). AVID: A global alignment program. *Genome Research*, 13(1), 97–102.
- Brennan, A. C., Méndez-Vigo, B., Haddioui, A., Martínez-Zapater, J. M., Picó, F. X., & Alonso-Blanco, C. (2014). The genetic structure of *Arabidopsis thaliana* in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. *BMC Plant Biology*, 14(1), 1–14.
- Brookes, A. J. (1999). The essence of SNPs. *Gene*, 234(2), 177–186.
- Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., & NISC Comparative Sequencing Program. (2003). LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13(4), 721–731.
- Bruskiewich, R. M., Cosico, A. B., Eusebio, W., Portugal, A. M., Ramos, L. M., Reyes, M. T., & McLaren, C. G. (2003). Linking genotype to phenotype: the international rice information system (IRIS). *Bioinformatics (Oxford, England)*, 19(suppl_1), i63–i65.
- Bundock, P. C., & Henry, R. J. (2004). Single nucleotide polymorphism, haplotype diversity and recombination in the *Isa* gene of barley. *Theoretical and Applied Genetics*, 109(3), 543–551.
- Burke, J., Davison, D., & Hide, W. (1999). d2_cluster: A validated method for clustering EST and full-length cDNA sequences. *Genome Research*, 9(11), 1135–1142.
- Buza, K., Wilczynski, B., & Dojer, N. (2015). RECORD: Reference-assisted genome assembly for closely related genomes. *International Journal of Genomics*, 2015.
- Byers, R. L., Harker, D. B., Yourstone, S. M., Maughan, P. J., & Udall, J. A. (2012). Development and mapping of SNP assays in allotetraploid cotton. *Theoretical and Applied Genetics*, 124(7), 1201–1214.
- Canzar, S., Andreotti, S., Weese, D., Reinert, K., & Klau, G. W. (2016). CIDANE: Comprehensive isoform discovery and abundance estimation. *Genome Biology*, 17(1), 1–18.
- Carpentier, S. C., Panis, B., Vertommen, A., Swennen, R., Sergeant, K., Renaut, J., & Devreese, B. (2008). Proteome analysis of non-model plants: A challenging but powerful approach. *Mass Spectrometry Reviews*, 27(4), 354–377.
- Carrillo-Perdomo, E., Vidal, A., Kreplak, J., Duborjal, H., Leveugle, M., Duarte, J., & Aubert, G. (2020). Development of new genetic resources for faba bean (*Vicia faba* L.) breeding through the discovery of gene-based SNP markers and the construction of a high-density consensus map. *Scientific Reports*, 10(1), 1–14.
- Challis, D., Yu, J., Evani, U. S., Jackson, A. R., Paithankar, S., Coarfa, C., & Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*, 13(1), 1–12.
- Chang, S., Puryear, J., & Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter*, 11(2), 113–116.
- Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., & Huang, X. (2015). Bridger: A new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology*, 16(1), 1–10.
- Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., & Mardis, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, 6(9), 677–681.
- Chiapparino, E., Lee, D., & Donini, P. (2004). Genotyping single nucleotide polymorphisms in barley by tetra-primer ARMS-PCR. *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, 47(2), 414–420.
- Chiara, M., Gioiosa, S., Chillemi, G., D'Antonio, M., Flati, T., Picardi, E., & Castrignanò, T. (2018). CoVaCS: A consensus variant calling system. *BMC Genomics*, 19(1), 1–9.
- Ching, A. D. A., & Rafalski, A. N. T. O. N. I. (2002). Rapid genetic mapping of ESTs using SNP pyrosequencing and indel analysis. *Cellular & Molecular Biology Letters*, 7(2B), 803–810.
- Cho, R. J., Mindrinos, M., Richards, D. R., Sapolsky, R. J., Anderson, M., Drenkard, E., & Oefner, P. J. (1999). Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nature Genetics*, 23(2), 203–207.
- Chong, X., Zhang, F., Wu, Y., Yang, X., Zhao, N., Wang, H., & Chen, F. (2016). A SNP-enabled assessment of genetic diversity, evolutionary relationships and the identification of candidate genes in chrysanthemum. *Genome Biology and Evolution*, 8(12), 3661–3671.
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2), 80–92.

- Clément, S., Fillon, J., Bousquet, J., & Beaulieu, J. (2010). TreeSNPs: a laboratory information management system (LIMS) dedicated to SNP discovery in trees. *Tree Genetics & Genomes*, 6(3), 435–438.
- Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., & Waugh, R. (2009). Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, 10(1), 1–13.
- Coles, N. D., Coleman, C. E., Christensen, S. A., Jellen, E. N., Stevens, M. R., Bonifacio, A., & Maughan, P. J. (2005). Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. *Plant Science*, 168(2), 439–447.
- Cortés, A. J., Chavarro, M. C., & Blair, M. W. (2011). SNP marker diversity in common bean (*Phaseolus vulgaris* L.). *Theoretical and Applied Genetics*, 123(5), 827.
- Delourme, R., Falentin, C., Fomeju, B. F., Boillot, M., Lassalle, G., André, I., & Pauquet, J. (2013). High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. *BMC Genomics*, 14(1), 1–18.
- Denti, L., Rizzi, R., Beretta, S., Della Vedova, G., Previtali, M., & Bonizzoni, P. (2018). ASGAL: Aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events. *BMC Bioinformatics*, 19(1), 1–21.
- Dereeper, A., Homa, F., Andres, G., Sempere, G., Sarah, G., Hueber, Y., & Ruiz, M. (2015). SNIPlay3: A web-based application for exploration and large scale analyses of genomic variations. *Nucleic Acids Research*, 43(W1), W295–W300.
- Dereeper, A., Nicolas, S., Le Cunff, L., Bacilieri, R., Doligez, A., Peros, J. P., & This, P. (2011). SNIPlay: A web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics*, 12(1), 1–14.
- Dobin, A., & Gingeras, T. R. (2016). *Optimizing RNA-Seq mapping with STAR. Data mining techniques for the life sciences* (pp. 245–262). New York, NY: Humana Press.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1), 15–21.
- Doyle, J.J. & Doyle, J.L. (1987). *A rapid DNA isolation procedure for small quantities of fresh leaf tissue* (No. RESEARCH).
- Droc, G., Perin, C., Fromentin, S., & Larmande, P. (2009). OryGenesDB 2008 update: Database interoperability for functional genomics of rice. *Nucleic Acids Research*, 37(suppl_1), D992–D995.
- Du, H., Yang, J., Chen, B., Zhang, X., Zhang, J., Yang, K., & Wen, C. (2019). Target sequencing reveals genetic diversity, population structure, core-SNP markers, and fruit shape-associated loci in pepper varieties. *BMC Plant Biology*, 19(1), 1–16.
- Duran, C., Appleby, N., Clark, T., Wood, D., Imelfort, M., Batley, J., & Edwards, D. (2009). AutoSNPdb: An annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Research*, 37(suppl_1), D951- distributed *Pinus sylvestris*. D953.
- Dvornyk, V., Sirviö, A., Mikkonen, M., & Savolainen, O. (2002). Low nucleotide diversity at the *pal1* locus in the widely. *Molecular Biology and Evolution*, 19(2), 179–188.
- Erwin, G. D., Oksenberg, N., Truty, R. M., Kostka, D., Murphy, K. K., Ahituv, N., & Capra, J. A. (2014). Integrating diverse datasets improves developmental enhancer prediction. *PLoS Computational Biology*, 10(6), e1003677.
- Feltus, F. A., Wan, J., Schulze, S. R., Estill, J. C., Jiang, N., & Paterson, A. H. (2004). An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Research*, 14(9), 1812–1819.
- Feng, J., Li, W., & Jiang, T. (2011). Inference of isoforms from short sequence reads. *Journal of Computational Biology*, 18(3), 305–321.
- Fong, C., Ko, D. C., Wasnick, M., Radey, M., Miller, S. I., & Brittnacher, M. (2010). GWAS analyzer: Integrating genotype, phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics (Oxford, England)*, 26(4), 560–564.
- Forconi, M., & Herschlag, D. (2009). Metal ion-based RNA cleavage as a structural probe. *Methods in Enzymology*, 468, 91–106.
- Foster, J. T., Allan, G. J., Chan, A. P., Rabinowicz, P. D., Ravel, J., Jackson, P. J., & Keim, P. (2010). Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). *BMC Plant Biology*, 10(1), 1–11.
- Fournier-Level, A., Wilczek, A. M., Cooper, M. D., Roe, J. L., Anderson, J., Eaton, D., & Schmitt, J. (2013). Paths to selection on life history loci in different natural environments across the native range of *Arabidopsis thaliana*. *Molecular Ecology*, 22(13), 3552–3566.
- Freudenthal, J. A., Ankenbrand, M. J., Grimm, D. G., & Korte, A. (2019). GWAS-Flow: A GPU accelerated framework for efficient permutation based genome-wide association studies. *BioRxiv*, 783100.
- Fu, Y. B., & Peterson, G. W. (2012). Developing genomic resources in two *Linum* species via 454 pyrosequencing and genomic reduction. *Molecular Ecology Resources*, 12(3), 492–500.
- Geleta, M., Gustafsson, C., Glaubitz, J. C., & Ortiz, R. (2020). High-density genetic linkage mapping of lepidium based on genotyping-by-sequencing SNPs and segregating contig tag haplotypes. *Frontiers in Plant Science*, 11, 448.
- Gontarz, P. M., Berger, J., & Wong, C. F. (2013). SRmapper: A fast and sensitive genome-hashing alignment tool. *Bioinformatics (Oxford, England)*, 29(3), 316–321.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644.
- Granata, I., Sangiovanni, M., Maiorano, F., Miele, M., & Guarracino, M. R. (2016). Var2GO: A web-based tool for gene variants selection. *BMC Bioinformatics*, 17(12), 376.
- Grivet, L., Glaszmann, J. C., Vincentz, M., Da Silva, F., & Arruda, P. (2003). ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes. *Theoretical and Applied Genetics*, 106(2), 190–197.
- Gupta, P. K., Roy, J. K., & Prasad, M. (2001). Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science*, 80(4), 524–535.

- Hach, F., Sarrafi, I., Hormozdiari, F., Alkan, C., Eichler, E. E., & Sahinalp, S. C. (2014). mrsFAST-Ultra: A compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Research*, *42*(W1), W494–W500.
- Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., . . . Buell, C. R. (2011). Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genomics*, *12*(1), 1–12.
- Han, Y., Kang, Y., Torres-Jerez, I., Cheung, F., Town, C. D., Zhao, P. X., & Monteros, M. J. (2011). Genome-wide SNP discovery in tetraploid alfalfa using 454 sequencing and high resolution melting analysis. *BMC Genomics*, *12*(1), 1–11.
- Hart, S. N., Duffy, P., Quest, D. J., Hossain, A., Meiners, M. A., & Kocher, J. P. (2016). VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. *Briefings in Bioinformatics*, *17*(2), 346–351.
- Hayashi, K., Hashimoto, N., Daigen, M., & Ashikawa, I. (2004). Development of PCR-based SNP markers for rice blast resistance genes at the Piz locus. *Theoretical and Applied Genetics*, *108*(7), 1212–1220.
- Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques*, *56*(2), 61–77.
- Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, *6*(2), 95–108.
- Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., & Hackermüller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational Biology*, *5*(9), e1000502.
- Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*(9), 868–877.
- Hyten, D. L., Cannon, S. B., Song, Q., Weeks, N., Fickus, E. W., Shoemaker, R. C., & Cregan, P. B. (2010). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics*, *11*(1), 1–8.
- Jacob, H. J., Lindpaintner, K., Lincoln, S. E., Kusumi, K., Bunker, R. K., Mao, Y. P., & Lander, E. S. (1991). Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. *Cell*, *67*(1), 213–224.
- Jammali, S., Aguilar, J. D., Kuitche, E., & Ouangraoua, A. (2019). SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups. *BMC Bioinformatics*, *20*(3), 37–52.
- Jander, G., Norris, S. R., Rounsley, S. D., Bush, D. F., Levin, I. M., & Last, R. L. (2002). *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiology*, *129*(2), 440–450.
- Jiang G.-L. (2013). Molecular markers and marker-assisted breeding in plants. In *Plant breeding from laboratories to fields*, IntechOpen, (pp 45–83). <<https://doi.org/10.5772/52583>>.
- Jones, E., Chu, W. C., Ayele, M., Ho, J., Bruggeman, E., Yourstone, K., & Warren, J. (2009). Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm. *Molecular Breeding*, *24*(2), 165–176.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006.
- Khiste, N., & Ilie, L. (2017). HISEA: Hierarchical seed aligner for pacbio data. *BMC Bioinformatics*, *18*(1), 1–13.
- Kielbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493.
- Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.
- Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915.
- Kim, K. D., Kang, Y., & Kim, C. (2020). Application of genomic big data in plant breeding: Past, present, and future. *Plants*, *9*(11), 1454.
- Kim, M. Y., Van, K., Lestari, P., Moon, J. K., & Lee, S. H. (2005). SNP identification and SNAP marker development for a GmNARK gene controlling supermodulation in soybean. *Theoretical and Applied Genetics*, *110*(6), 1003–1010.
- Kim, S. R., Ramos, J., Ashikari, M., Virk, P. S., Torres, E. A., Nissila, E., & Jena, K. K. (2016). Development and validation of allele-specific SNP/indel markers for eight yield-enhancing genes using whole-genome sequencing strategy to increase yield potential of rice, *Oryza sativa* L. *Rice*, *9*(1), 1–17.
- Klus, P., & Lam, S. (2012). Dag Lyberg, Ming Sin Cheung, Graham Pullan, Ian McFarlane, Giles SH Yeo, and Brian YH Lam. BarraCUDA-a fast short read sequence aligner using graphics processing units. *BMC Research Notes*, *5*(1), 1–7.
- Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., & Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England)*, *25*(17), 2283–2285.
- Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Aguilera, M. A., Meyer, R., & Massouras, A. (2019). VarSome: The human genomic variant search engine. *Bioinformatics (Oxford, England)*, *35*(11), 1978.
- Kota, R., Varshney, R. K., Thiel, T., Dehmer, K. J., & Graner, A. (2001). Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). *Hereditas*, *135*(2-3), 145–151.
- Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics*, *17*(1), 21–24.
- Kumar, S., Agarwal, S., & Ranvijay. (2019). Fast and memory efficient approach for mapping NGS reads to a reference genome. *Journal of Bioinformatics and Computational Biology*, *17*(02), 1950008.
- Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., & Dry, J. R. (2016). VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, *44*(11), e108–e108.
- Lander, E. S., & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, *121*(1), 185–199.

- Landjeva, S., Korzun, V., & Börner, A. (2007). Molecular markers: Actual and potential contributions to wheat genome characterization and breeding. *Euphytica*, 156(3), 271–296.
- Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics*, 32(1), 11–17.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357–359.
- Lateef, D. D. (2015). DNA marker technologies in plants and applications for crop improvements. *Journal of Biosciences and Medicines*, 3(05), 7.
- Lazzari, B., Caprera, A., Vecchiotti, A., Merelli, I., Barale, F., Milanese, L., & Pozzi, C. (2008). Version VI of the ESTree db: An improved tool for peach transcriptome analysis. *BMC Bioinformatics*, 9(2), 1–6.
- Lazzari, B., Caprera, A., Vecchiotti, A., Stella, A., Milanese, L., & Pozzi, C. (2005). ESTree db: A tool for peach functional genomics. *BMC Bioinformatics*, 6(4), 1–6.
- Le Dantec, L., Chagne, D., Pot, D., Cantin, O., Garnier-Gere, P., Bedon, F., & Plomion, C. (2004). Automated SNP detection in expressed sequence tags: Statistical considerations and application to maritime pine sequences. *Plant Molecular Biology*, 54(3), 461–470.
- Le, S. Q., & Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, 21(6), 952–960.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21), 2987–2993.
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, 34(18), 3094–3100.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), 1754–1760.
- Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851–1858.
- Li, M. X., Gui, H. S., Kwan, J. S., Bao, S. Y., & Sham, P. C. (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Research*, 40(7), e53–e53.
- Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: Short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, 24(5), 713–714.
- Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., & Wang, J. (2009). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England)*, 25(15), 1966–1967.
- Li, W., Feng, J., & Jiang, T. (2011). IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *Journal of Computational Biology*, 18(11), 1693–1707.
- Lin, H. N., & Hsu, W. L. (2017). Kart: A divide-and-conquer algorithm for NGS read alignment. *Bioinformatics (Oxford, England)*, 33(15), 2281–2287.
- Lin, H. N., & Hsu, W. L. (2020). GSAAlign: an efficient sequence alignment tool for intra-species genomes. *BMC genomics*, 21(1), 182. Available from <https://doi.org/10.1186/s12864-020-6569-1>.
- Lippert, R. A. (2005). Space-efficient whole genome comparisons with Burrows–Wheeler transforms. *Journal of Computational Biology*, 12(4), 407–415.
- Liu, J., Li, G., Chang, Z., Yu, T., Liu, B., McMullen, R., & Huang, X. (2016). BinPacker: Packing-based de novo transcriptome assembly from RNA-seq data. *PLoS Computational Biology*, 12(2), e1004772.
- Liu, J., Yu, T., Jiang, T., & Li, G. (2016). TransComb: Genome-guided transcriptome assembly via combing junctions in splicing graphs. *Genome Biology*, 17(1), 1–9.
- Liu, J., Yu, T., Mu, Z., & Li, G. (2019). TransLiG: A de novo transcriptome assembler that uses line graph iteration. *Genome Biology*, 20(1), 1–9.
- Liu, W., Wu, S., Lin, Q., Gao, S., Ding, F., Zhang, X., & Hu, S. (2018). RGAAT: A reference-based genome assembly and annotation tool for new genomes and upgrade of known genomes. *Genomics, Proteomics & Bioinformatics*, 16(5), 373–381.
- Liu, X., Jian, X., & Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human mutation*, 32(8), 894–899.
- Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation*, 37(3), 235–241.
- Livak, K. J. (1999). Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. *Genetic Analysis: Biomolecular Engineering*, 14(5–6), 143–149.
- Livak, K. J., Flood, S. J., Marmaro, J., Giusti, W., & Deetz, K. (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *Genome Research*, 4(6), 357–362.
- Lopez, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J., & Verdier, V. (2005). Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). *Theoretical and Applied Genetics*, 110(3), 425–431.
- Luo, H., Zhao, W., Wang, Y., Xia, Y., Wu, X., Zhang, L., & Jing, H. C. (2016). SorGSD: A sorghum genome SNP database. *Biotechnology for Biofuels*, 9(1), 1–9.
- Mackill, D. J., Nguyen, H. T., & Zhang, J. (1999). Use of molecular markers in plant improvement programs for rainfed lowland rice. *Field Crops Research*, 64(1–2), 177–185.
- Makarov, V., O’Grady, T., Cai, G., Lihm, J., Buxbaum, J. D., & Yoon, S. (2012). AnnTools: A comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics (Oxford, England)*, 28(5), 724–725.

- Malhis, N., & Jones, S. J. (2010). High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics (Oxford, England)*, 26(8), 1029–1035.
- Manske, H. M., & Kwiatkowski, D. P. (2009). SNP-o-matic. *Bioinformatics (Oxford, England)*, 25(18), 2434–2435.
- Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, 14(1), e1005944.
- Maretty, L., Sibbesen, J. A., & Krogh, A. (2014). Bayesian transcriptome assembly. *Genome Biology*, 15(10), 1–11.
- Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S., & Morris, R. W. (2010). SeqEM: An adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics (Oxford, England)*, 26(22), 2803–2810.
- Matthews, D. E., Carollo, V. L., Lazo, G. R., & Anderson, O. D. (2003). GrainGenes, the genome database for small-grain crops. *Nucleic Acids Research*, 31(1), 183–186.
- Matukumalli, L. K., Grefenstette, J. J., Hyten, D. L., Choi, I. Y., Cregan, P. B., & Van Tassell, C. P. (2006). SNP-PHAGE—High throughput SNP discovery pipeline. *BMC Bioinformatics*, 7(1), 1–7.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303.
- McNally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., & Leach, J. E. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences*, 106(30), 12273–12278.
- Melo, A. T., Bartaula, R., & Hale, I. (2016). GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics*, 17(1), 1–15.
- Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews. Genetics*, 11(1), 31–46.
- Mezlini, A. M., Smith, E. J., Fiume, M., Buske, O., Savich, G. L., Shah, S., & Brudno, M. (2013). iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Research*, 23(3), 519–529.
- Mitha, F., Herodotou, H., Borisov, N., Jiang, C., Yoder, J., & Owzar, K. (2011). SNPpy-Database management for SNP data from GWAS studies. *Duke Biostatistics and Bioinformatics (B&B) Working Paper Series*, 14.
- Morales, M., Roig, E., Monforte, A. J., Arus, P., & Garcia-Mas, J. (2004). Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, 47(2), 352–360.
- Morgulis, A., & Agarwala, R. (2020). SRPRISM (Single Read Paired Read Indel Substitution Minimizer): An efficient aligner for assemblies with explicit guarantees. *GigaScience*, 9(4), g1aa023.
- Nakato, R., & Gotoh, O. (2010). Cgaln: fast and space-efficient whole-genome alignment. *BMC Bioinformatics*, 11(1), 1–14.
- Nelson, J. C., Wang, S., Wu, Y., Li, X., Antony, G., White, F. F., & Yu, J. (2011). Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. *BMC Genomics*, 12(1), 352.
- Ning, Z., Cox, A. J., & Mullikin, J. C. (2001). SSAHA: A fast search method for large DNA databases. *Genome Research*, 11(10), 1725–1729.
- Oliver, R. E., Lazo, G. R., Lutz, J. D., Rubenfield, M. J., Tinker, N. A., Anderson, J. M., ... Jackson, E. W. (2011). Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. *BMC Genomics*, 12(1), 1–15.
- Orro, A., Guffanti, G., Salvi, E., Macciardi, F., & Milanese, L. (2008). SNPLims: A data management system for genome wide association studies. *BMC Bioinformatics*, 9(S2), S13.
- Ortiz, R. (2010). Molecular plant breeding. *Crop Science*, 50(5), 2196.
- Osman, A., Jordan, B., Lessard, P. A., Muhammad, N., Haron, M. R., Riffin, N. M., & Housman, D. E. (2003). Genetic diversity of *Eurycoma longifolia* inferred from single nucleotide polymorphisms. *Plant Physiology*, 131(3), 1294–1301.
- Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., & Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Research*, 18(12), 2024–2033.
- Paila, U., Chapman, B. A., Kirchner, R., & Quinlan, A. R. (2013). GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Computational Biology*, 9(7), e1003153.
- Paris, M., Jones, M. G., & Eglinton, J. K. (2002). Genotyping single nucleotide polymorphisms for selection of barley β -amylase alleles. *Plant Molecular Biology Reporter*, 20(2), 149–159.
- Park, D. J., Nguyen-Dumont, T., Kang, S., Verspoor, K., & Pope, B. J. (2014). Annokey: An annotation tool based on key term search of the NCBI Entrez Gene database. *Source Code for Biology and Medicine*, 9(1), 15.
- Paule, J., Heller, S., Maciel, J. R., Monteiro, R. F., Leme, E. M., & Zizka, G. (2020). Early diverging and core Bromelioideae (*Bromeliaceae*) reveal contrasting patterns of genome size evolution and polyploidy. *Frontiers in Plant Science*, 11.
- Peng, Y., Leung, H. C., Yiu, S. M., Lv, M. J., Zhu, X. G., & Chin, F. Y. (2013). IDBA-tran: A more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics (Oxford, England)*, 29(13), i326–i334.
- Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., & Tsai, J. (2003). TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics (Oxford, England)*, 19(5), 651–652.
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33, 290–295.
- Petersen, J. L., & Coleman, S. J. (2020). Next-generation sequencing in equine genomics. *Veterinary Clinics: Equine Practice*, 36(2), 195–209.
- Plake, C., Royer, L., Winnenburger, R., Hakenberg, J., & Schroeder, M. (2009). GoGene: Gene annotation in the fast lane. *Nucleic Acids Research*, 37 (suppl_2), W300–W304.
- Powell, W., Machray, G. C., & Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends in Plant Science*, 1(7), 215–222.

- Qi, L., & Ma, G. (2020). Marker-assisted gene pyramiding and the reliability of using SNP markers located in the recombination suppressed regions of sunflower (*Helianthus annuus* L.). *Genes*, *11*(1), 10.
- Qi, L. L., Talukder, Z. I., Hulke, B. S., & Foley, M. E. (2017). Development and dissection of diagnostic SNP markers for the downy mildew resistance genes *Pl Arg* and *Pl 8* and maker-assisted gene pyramiding in sunflower (*Helianthus annuus* L.). *Molecular Genetics and Genomics*, *292*(3), 551–563.
- Qi, P., Gimode, D., Saha, D., Schröder, S., Chakraborty, D., Wang, X., & Devos, K. M. (2018). UGbS-Flex, a novel bioinformatics pipeline for imputation-free SNP discovery in polyploids without a reference genome: Finger millet as a case study. *BMC Plant Biology*, *18*(1), 1–19.
- Rafalski, J. A. (2002). Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science*, *162*(3), 329–333.
- Rebbeck, T. R., Spitz, M., & Wu, X. (2004). Assessing the function of genetic variants in candidate gene association studies. *Nature Reviews. Genetics*, *5*(8), 589–597.
- Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S., Consortium., ... Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, *46*(8), 912–918. Available from <https://doi.org/10.1038/ng.3036>.
- Rio, D. C., Ares, M., Hannon, G. J., & Nilsen, T. W. (2010). Purification of RNA using TRIzol (TRI reagent). *Cold Spring Harbor Protocols*, *2010* (6), pdb-prot5439.
- Rudd, S., Schoof, H., & Mayer, K. (2005). PlantMarkers—A database of predicted molecular markers from plants. *Nucleic Acids Research*, *33* (suppl_1), D628–D632.
- Russell, J., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G., & Powell, W. (2004). A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, *47*(2), 389–398.
- Salgotra, R. K., & Stewart, C. N. (2020). Functional markers for precision plant breeding. *International Journal of Molecular Sciences*, *21*(13), 4792.
- Sanchez-Villeda, H., Schroeder, S., Polacco, M., McMullen, M., Havermann, S., Davis, G., & Schultz, L. (2003). Development of an integrated laboratory information management system for the maize mapping project. *Bioinformatics (Oxford, England)*, *19*(16), 2022–2030.
- Scheben, A., Verpaalen, B., Lawley, C. T., Chan, C. K. K., Bayer, P. E., Batley, J., & Edwards, D. (2019). CropSNPdb: A database of SNP array data for Brassica crops and hexaploid bread wheat. *The Plant Journal*, *98*(1), 142–152.
- Schmid, K. J., Sørensen, T. R., Stracke, R., Törjék, O., Altmann, T., Mitchell-Olds, T., & Weisshaar, B. (2003). Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Research*, *13*(6a), 1250–1257.
- Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, *28*(8), 1086–1092.
- Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., & Miller, W. (2003). Human–mouse alignments with BLASTZ. *Genome Research*, *13*(1), 103–107.
- Shao, M., & Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*, *35*(12), 1167–1169.
- Shavrukov, Y., Suchecki, R., Eliby, S., Abugalieva, A., Kenebayev, S., & Langridge, P. (2014). Application of next-generation sequencing technology to study genetic diversity and identify unique SNP markers in bread wheat from Kazakhstan. *BMC Plant Biology*, *14*(1), 1–13.
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123.
- Singh, N., Choudhury, D. R., Singh, A. K., Kumar, S., Srinivasan, K., Tyagi, R. K., ... Singh, R. (2013). Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS One*, *8*(12), e84136.
- Singh, T. R., Gupta, A., Riju, A., Mahalaxmi, M., Seal, A., & Arunachalam, V. (2013). Computational identification and analysis of single-nucleotide polymorphisms and insertions/deletions in expressed sequence tag data of Eucalyptus. *Journal of Genetics*, *92*(2), 34–38.
- Souaiaia, T., Frazier, Z., & Chen, T. (2011). ComB: SNP calling and mapping analysis for color and nucleotide space platforms. *Journal of Computational Biology*, *18*(6), 795–807.
- Suarez, H. G., Langer, B. E., Ladde, P., & Hiller, M. (2017). chainCleaner improves genome alignment specificity and sensitivity. *Bioinformatics (Oxford, England)*, *33*(11), 1596–1603.
- Surget-Groba, Y., & Montoya-Burgos, J. I. (2010). Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research*, *20*(10), 1432–1440.
- Suwarno, W. B., Pixley, K. V., Palacios-Rojas, N., Kaeppeler, S. M., & Babu, R. (2015). Genome-wide association analysis reveals new targets for carotenoid biofortification in maize. *Theoretical and Applied Genetics*, *128*(5), 851–864.
- Swidan, F., Rocha, E. P., Shmoish, M., & Pinter, R. Y. (2006). An integrative method for accurate comparative genome mapping. *PLoS Computational Biology*, *2*(8), e75.
- Syvänen, A. C., Aalto-Setälä, K., Harju, L., Kontula, K., & Söderlund, H. (1990). A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics*, *8*(4), 684–692.
- Szkiba, D., Kapun, M., von Haeseler, A., & Gallach, M. (2014). SNP2GO: Functional analysis of genome-wide association studies. *Genetics*, *197*(1), 285–289.
- Tang, J., Leunissen, J. A., Voorrips, R. E., van der Linden, C. G., & Vosman, B. (2008). HaploSNPer: A web-based allele and SNP detection tool. *BMC Genetics*, *9*(1), 1–7.
- Tang, J., Vosman, B., Voorrips, R. E., van der Linden, C. G., & Leunissen, J. A. (2006). QualitySNP: A pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics*, *7*(1), 1–15.

- Tárraga, J., Arnau, V., Martínez, H., Moreno, R., Cazorla, D., Salavert-Torres, J., & Medina, I. (2014). Acceleration of short and long DNA read mapping without loss of accuracy using suffix array. *Bioinformatics (Oxford, England)*, *30*(23), 3396–3398.
- Tenaillon, M. I., Sawkins, M. C., Anderson, L. K., Stack, S. M., Doebley, J., & Gaut, B. S. (2002). Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Genetics*, *162*(3), 1401–1413.
- Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation*, *32*(4), 358–368.
- Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., & Belzile, F. (2017). Fast-GBS: A new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics*, *18*(1), 1–7.
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, *25*(9), 1105–1111. Available from <https://doi.org/10.1093/bioinformatics/btp120>.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.
- Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, *15*(11), 1–15.
- Trebbi, D., Maccaferri, M., de Heer, P., Sørensen, A., Giuliani, S., Salvi, S., & Tuberosa, R. (2011). High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf.). *Theoretical and Applied Genetics*, *123*(4), 555–569.
- Trebbi, D., Ravi, S., Broccanello, C., Chiodi, C., Francis, G., Oliver, J., & Stevanato, P. (2019). Identification and validation of SNP markers linked to seed toxicity in *Jatropha curcas* L. *Scientific Reports*, *9*(1), 1–7.
- Tyrka, M., Blaszczyk, L., Chelkowski, J., Lind, V., Kramer, I., Weilepp, M., & Ordon, F. R. A. N. K. (2004). Development of the single nucleotide polymorphism marker of the wheat Lr1 leaf rust resistance gene. *Cellular & Molecular Biology Letters*, *9*(4B), 879–889.
- Van Inghelandt, D., Melchinger, A. E., Lebreton, C., & Stich, B. (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theoretical and Applied Genetics*, *120*(7), 1289–1299.
- Vigouroux, Y., Mitchell, S., Matsuoka, Y., Hamblin, M., Kresovich, S., Smith, J. S. C., & Doebley, J. (2005). An analysis of genetic diversity across the maize genome using microsatellites. *Genetics*, *169*(3), 1617–1630.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. V. D., Hornes, M., & Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, *23*(21), 4407–4414.
- Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T., & Itakura, K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to Φ X 174 DNA: The effect of single base pair mismatch. *Nucleic Acids Research*, *6*(11), 3543–3558.
- Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., & Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science (New York, N.Y.)*, *280*(5366), 1077–1082.
- Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164.
- Waugh, R., Jannink, J. L., Muehlbauer, G. J., & Ramsay, L. (2009). The emergence of whole genome association scans in barley. *Current Opinion in Plant Biology*, *12*(2), 218–222.
- Weckwerth, W., Ghatak, A., Bellaire, A., Chaturvedi, P., & Varshney, R. K. (2020). PANOMICS meets germplasm. *Plant Biotechnology Journal*, *18*(7), 1507–1525.
- Weese, D., Holtgrewe, M., & Reinert, K. (2012). RazerS 3: Faster, fully sensitive read mapping. *Bioinformatics (Oxford, England)*, *28*(20), 2592–2599.
- Wei, Z., Wang, W., Hu, P., Lyon, G. J., & Hakonarson, H. (2011). SNVer: A statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*, *39*(19), e132–e132.
- Welsh, J., & McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, *18*(24), 7213–7218.
- Wheeler, S. J., Church, D. M., & Ostell, J. M. (2001). Spidey: A tool for mRNA-to-genomic alignments. *Genome Research*, *11*(11), 1952–1957.
- Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*, *18*(1), 1–12.
- Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, *18*(22), 6531–6535.
- Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., & Nagarajan, N. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201.
- Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*(7), 873–881.
- Wünsch, C., Banck, H., Müller-Tidow, C., & Dugas, M. (2020). AMLVaran: A software approach to implement variant analysis of targeted NGS sequencing data in an oncological care setting. *BMC Medical Genomics*, *13*(1), 17.
- Xia, W., Luo, T., Zhang, W., Mason, A. S., Huang, D., Huang, X., & Xiao, Y. (2019). Development of high-density SNP markers and their application in evaluating genetic diversity and population structure in *Elaeis guineensis*. *Frontiers in Plant Science*, *10*, 130.
- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., & Wang, J. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics (Oxford, England)*, *30*(12), 1660–1666.
- Xing, Q., Ru, Z., Li, J., Zhou, C., Jin, D., Sun, Y., & Wang, B. (2005). Cloning a second form of adenine phosphoribosyl transferase gene (TaAPT2) from wheat and analysis of its association with thermo-sensitive genic male sterility (TGMS). *Plant Science*, *169*(1), 37–45.

- Xu, Y., Li, P., Zou, C., Lu, Y., Xie, C., Zhang, X., & Olsen, M. S. (2017). Enhancing genetic gain in the era of molecular breeding. *Journal of Experimental Botany*, *68*(11), 2641–2666.
- Yamamoto, T., Nagasaki, H., Yonemaru, J. I., Ebana, K., Nakajima, M., Shibaya, T., & Yano, M. (2010). Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. *BMC Genomics*, *11*(1), 1–14.
- You, F. M., Huo, N., Deal, K. R., Gu, Y. Q., Luo, M. C., McGuire, P. E., & Anderson, O. D. (2011). Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics*, *12*(59). Available from <https://doi.org/10.1186/1471-2164-12-59>.
- Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., ... Zhang, Q. (2011). Gains in QTL detection using an ultra-high density SNP using next-generation sequencing without a reference genome sequence. *BMC Genomics*, *12*(1), 59. map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS one*, *6*(3), e17595.
- Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*, *31*(1), 11–15.
- Zhang, X., & Borevitz, J. O. (2009). Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics*, *182*(4), 943–954.
- Zhang, Z., Guo, X., Liu, B., Tang, L., & Chen, F. (2011). Genetic diversity and genetic relationship of *Jatropha curcas* between China and Southeast Asian revealed by amplified fragment length polymorphisms. *African Journal of Biotechnology*, *10*(15), 2825–2832.
- Zhao, W., Canaran, P., Jurkuta, R., Fulton, T., Glaubitz, J., Buckler, E., & Kresovich, S. (2006). Panzea: A database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Research*, *34*(suppl_1), D752–D757.
- Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., & Zhang, Y. (2004). BGI-RIS: An integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Research*, *32*(suppl_1), D377–D382.
- Zhu, Y. L., Song, Q. J., Hyten, D. L., Van Tassell, C. P., Matukumalli, L. K., Grimm, D. R., ... Cregan, P. B. (2003). Single-nucleotide polymorphisms in soybean. *Genetics*, *163*(3), 1123–1134.

Bioinformatics intervention in identification and development of molecular markers: an overview

Vikas Dwivedi^{1,2}, Lalita Pal² and Dinesh Yadav³

¹Agricultural Research Organization, The Volcani Center, Rishon LeZion, Israel, ²National Institute of Plant Genome Research, New Delhi, New Delhi, India, ³Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, Uttar Pradesh, India

32.1 Introduction

Due to change in their cellular and environmental factors, all organisms have mutations in the genome responsible for genetic variations and thus show polymorphism. If these mutations occur within a nucleotide sequence of a gene, the whole amino acid string of an open reading frame changed and form a functionally different new variant. Markers are “character traits” whose patterns of inheritance can be traced in plants at different levels. The markers are used to get more information regarding the genetics of other interested traits. Markers can be divided into three groups: morphological markers (i.e., seed shape, seed coat color, and flower color), cytological markers (chromosome karyotypes, bandings, and repeats), and biochemical markers (isozymes) (Nadeem, Nawaz, Shahid, Doğan, & Comertpay, 2018).

In the early 20th century, scientists found genes that are arranged in linear order on chromosomes. Genes can be linked and could be inherited in a group. The flanking gene within an area of the defined close interval are called molecular DNA markers. Molecular markers, gene or identifiable DNA sequence, found in the genome at fixed locations and show association with the trait.

A genetic marker may be mini and microsatellites (long DNA sequence) or single-nucleotide polymorphism, SNP (short DNA sequence)

Molecular markers are the source of potent informational tools that can be used to divulge the genetic exclusivity of individuals, species, and populations in plant breeding (Davey, Hohenlohe, & Etter, 2011).

There are more molecular markers for the classification of genotype. Morphological traits are affected by the environment but not molecular markers (Staub, Serquen, & Mccreight, 1997). In the breeding program, molecular marker data would be of great interest to investigate the correlation between phenotypically and genetically similar cultivars (Duzyaman, 2005). For crop improvement, different molecular techniques can be used to detect differences in between DNA of separate plants (Jonah, Bello, Lucky, & A. Midau, 2011). The molecular markers can be a “sign posts” and help to identify genes of interest. Molecular markers can be used in marker-assisted selection (MAS; Hoisington et al., 2002). The information of quantitative trait loci (QTL) can dissect by markers.

The discovery of polymerase chain reaction (PCR) was brought a new class of DNA markers, which can be used in the cloning of important genes by map-based cloning, synteny mapping, to get desirable genotypes by MAS, variability studies, and phylogenetic analysis (Joshi & Deshpande, 2011).

To analyze different objective, DNA markers show many advantages over traditional phenotypic markers as they provide more data. Therefore there are several types of molecular marker in plants, and each of them has their own advantages and disadvantages (Cadalen et al., 1998). The development of molecular markers (or DNA) has changed a lot in plant genetics and widely used in the area of the breeding program to improve varieties (Collard & Mackill, 2008).

32.2 Genetic markers

Genetic markers have an important role in the plant-breeding field. Moreover, genetic markers are act as flags or sign; linked with the target gene (Kebriyae, Kordrostami, & Rezadoost, 2012). Genetic markers are mainly divided into two types: (1) classical and (2) DNA/molecular markers.

32.2.1 Classical markers: The classical markers are further divided that include morphological markers, cytological markers and biochemical markers

32.2.1.1 Morphological markers

Morphological markers could be used to discriminate qualities like growth pattern, color of flower, seed structure, and other important agronomic traits by simply visualizing these characters. These markers are not required specific instruments and easy to use. They required simple biochemical and molecular techniques. Breeders used successfully these markers in the different breeding plants for numerous crops. However, there are some demerits of these markers likewise: the plant growth stages influenced these markers. They are limited in number and affected by various environmental factors (Eagles, Bariana, & Ogbonnaya, 2001). Humans used various morphological markers in plant breeding to develop new varieties since ancient times (Karaköy, Baloch, & Toklu, 2014).

32.2.1.2 Cytological markers

In cytology, chromosome karyotype and bands can show the structural features of chromosomes. The cytological markers are related with different banding patterns. They also changed according to size, shape, order, numbers, and position of chromosomes. These variations can relate to distributions of euchromatin and heterochromatin in chromosomes. There are different types of bands like G bands (Giemsa stain), Q bands (quinacrine hydrochloride), and R bands (inverted G bands). These signs could be used in the characterization of normal and altered chromosomes. These markers helped in the identification of linkage groups and also in the physical mapping (Jiang, 2013). The physical maps made by morphological and cytological markers can be used for genetic linkage mapping with the help of molecular biology techniques. Nevertheless, there are only a few reports that show direct use of cytological markers in genetic mapping and plant breeding.

32.2.1.3 Biochemical markers

Protein markers can be cataloged into markers. Isozymes are used as biochemical markers. These are different molecular forms or structural variants of enzymes. These are coded by various genes and have different molecular weights. As they also count for difference in electrophoretic mobility but show the same functions. They are allelic variations of the same genes. The electrophoretic mobility shift is due to amino acid substitution (point mutation) (Xu, 2010). Biochemical markers are used in the detection of population structure, subdivision of population, gene flow, and genetic diversity (Mateu-Andres & De Paco, 2005). They are codominant, cost-effective, and easily available. They can also be used in seed purity and sporadically in plant breeding due to less number.

32.2.2 Molecular markers

Molecular markers are used for the analysis of genetic variation among individuals as they easily link the phenotypic variation with genotypic variation. These have been used in the agronomic sector in recent decades. These are nucleotide sequences. The nucleotide sequences are different between individuals and showed polymorphism that leads to develop a marker. There are different types of molecular markers. During the last few decades, different systems continuously evolved like in the 1980s restriction fragment length polymorphisms (RFLPs) were prominent but nowadays' high-throughput sequencing-based SNPs are prominent markers (Fig. 32.1)

32.3 Restriction fragment length polymorphism (RFLP)

RFLPs, hybridization-based, are the most studied molecular markers among the different markers. It was first discovered and was used in human genome mapping. The polymorphisms in RFLPs based on the deletion, insertion, point mutation, or transposons. First, restriction enzymes digest the DNA, run on agarose gel, and then transfer to nylon membranes, followed by the hybridization of probes. The polymorphisms are restriction fragments of different sizes.

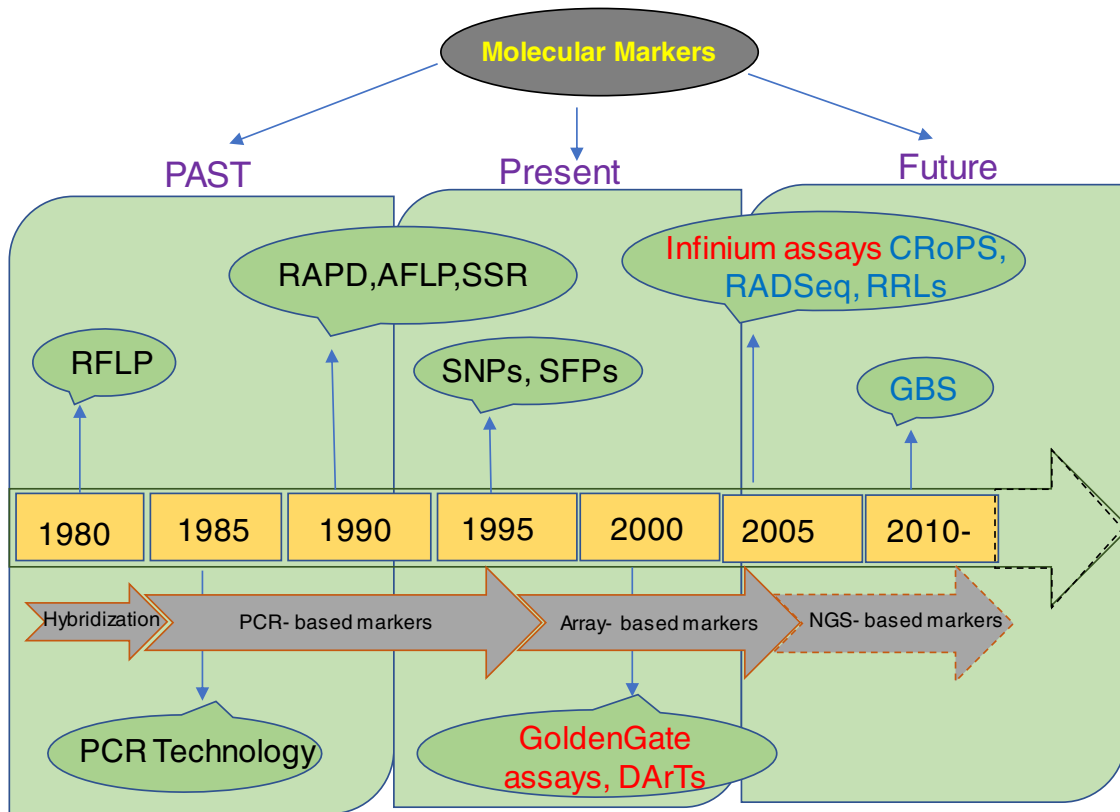


FIGURE 32.1 Schematic representation of molecular markers in the 1980s RFLPs (hybridization) but now SNPs markers based on new sequencing technologies. *RFLP*, restriction fragment length polymorphism; *SNP*, single-nucleotide polymorphism.

The size depends on the type of restriction enzymes. Double-stranded DNA cleaved by restriction enzymes at specific recognition sites (short sequences of DNA that are interspersed in plant and animal DNA) (Voet & Voet, 2004). RFLPs are used for construction of genetic maps. They are used for the analysis of genomics for comparison. RFLPs are mostly time-consuming and labor-dependent. For some plants, RFLPs show less polymorphisms. In many living organisms, insertion or deletion can take place within restriction sites. Sometimes difference of restriction sites created due to change in base pair. This difference resulting in base pair changes. Due to these variations, the recognition sites may alternate or eliminate. Therefore, when homologous chromosomes are imperiled to digestion, different products are created that can be identified by agarose gel and DNA hybridization.

RFLP markers are a powerful tool used for comparative as well as synteny mapping. Most RFLP markers are locus-specific and codominant in nature. It is a simple method as no distinct apparatus is required. In RFLP, there is no need to know the sequence used for a probe. There is only one need that is a genomic clone for the polymorphism. There are very few RFLPs that have been sequenced. It uses larger amounts of DNA. It is very difficult to automate. In 1980s and 1990s it was predominant, but since last decade, it lost its faith in breeding. Most plant breeders face the problem as it is expensive and requires quality DNA for the study.

Soller and Beckmann (1983) were the first persons to analyze the roles of RFLP in plants. They correlated it with varietal identification, marker-assisted introgression, surveys of genetic polymorphism, and identification and mapping of quantitative traits. There are some benefits of RFLP markers in genetic analysis that included multiple allelic forms, no pleiotropic effects on different traits, and lack of dominance (codominance).

32.3.1 Application of restriction fragment length polymorphism

32.3.1.1 Restriction fragment length polymorphism in DNA fingerprinting

RFLP is used in DNA fingerprinting. The restriction fragments pattern showed that this is the fingerprint for a clone. It can also be used in the identification of an individual plant and cultivar. A single probe gives a little information; therefore to yield fingerprints, information from numerous probes should be combined. Scientists have used the probe

approach for rice (Wang & Tanksley, 1989). The bacteriophage probe M13 has been used in fingerprinting of plants, both angiosperms and gymnosperms (Nybom, Rogstad, & Schaal, 1990).

32.3.1.2 Restriction fragment length polymorphism in species identification

RFLP data from chloroplast have been widely used in the identification of plant. Chloroplast genomes are more conserved and smaller. Complete sets of probes are easily accessible. Many scientists have used the chloroplast RFLP data in several plants species to identify and authenticate varieties and species like in *Phyllanthus* species (Sarin, Mohanty, & demente, 2013)

32.3.1.3 Restriction fragment length polymorphism in comparative mapping

RFLP maps can be gathered valuable data on taxonomic associations and chromosome progression through the comparison of maps. The tomato RFLP map has transferred to two Solanaceae species: potato and pepper by Tanksley, Bernatzky, Lapitan, and Prince, 1988.

32.3.1.4 Linkage mapping with restriction fragment length polymorphism markers

Usually, RFLP maps are used for self-pollinating plants or plants that can produce inbred lines by self-fertilization. All loci in these plants are generally homozygous. The inheritance in RFLP is different from the conventional one (Fig. 32.2). The RFLP markers help in plant breeding to find the tight linkage between markers and genes of interest. Such linkage gives information about desirable gene with help of an RFLP marker. Breeders frequently transferred

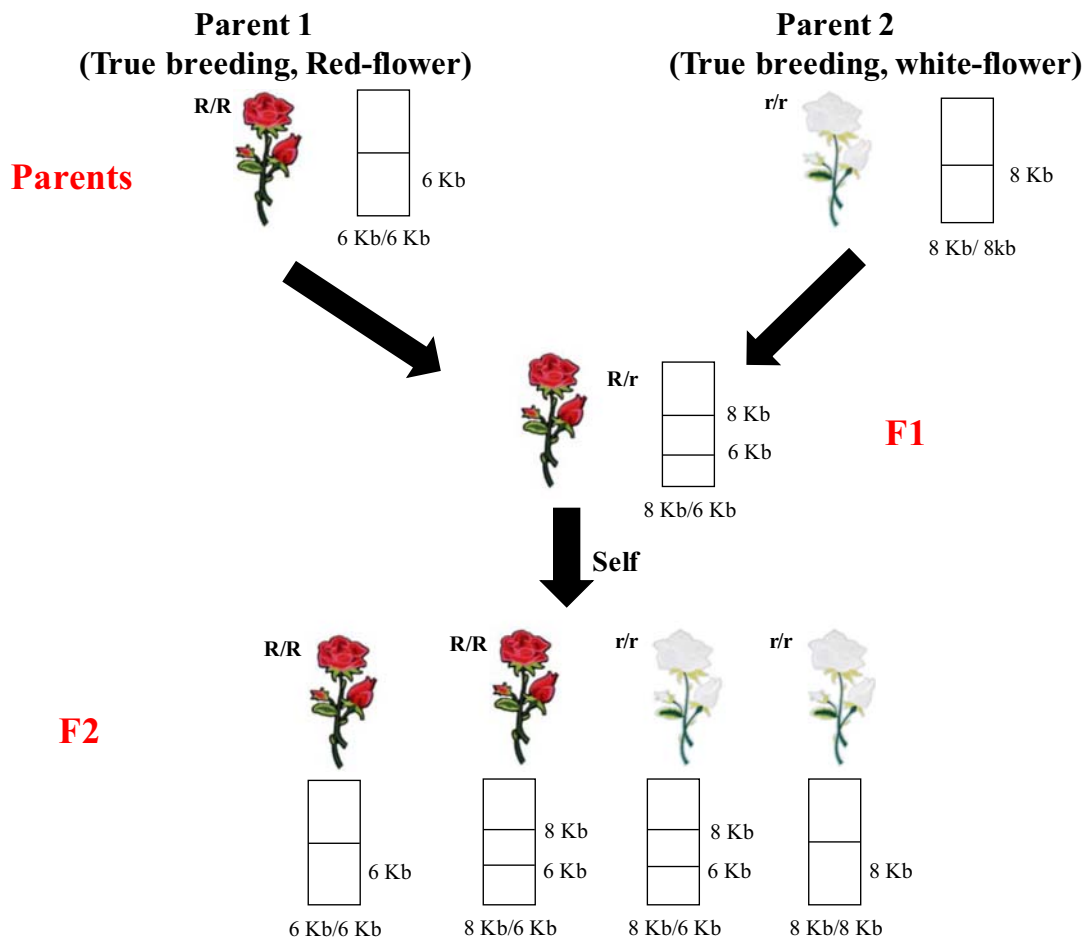


FIGURE 32.2 Comparison between RFLP marker and conventional marker controlling color of flower. Here the conserved gene is dominant. The plant is homozygous at all loci. The RFLP marker showed a segregation ratio of 1:2:1 in F2 population. *RFLP*, restriction fragment length polymorphism.

single-gene traits from one genetic background to other background. There is a problem to find the disease-resistance varieties from a population by the conventional method. Traditionally to find the variety with disease resistance, plants should be inoculated by pathogens. There is also a problem to screen the plants for different pathogens simultaneously. In contrast, detecting plants with resistance gene by their linkage to RFLP probe makes it easy without the need to inoculate the population with pathogens.

32.3.1.5 *Elucidating the genetic traits*

Many heritable traits are regulated by many genes together. Such characteristics are called quantitative traits as the amount of seeds, drought resistance. High-density RFLP maps help to measure the effect of genes for complex traits. RFLP markers have an ability to detect QTL based on the size of population being studied, effect of QTLs on character, and recombination frequency between marker and QTL. Complete RFLP maps help to detect QTLs for traits.

32.3.1.6 *Restriction fragment length polymorphism in back crossing*

The cross between donor parent and progeny is called back cross. This is developed for the recovery of the desired traits from genotypes. The selected progenies are crossed again to one genotype. It is also used to mark QTLs and genes in the genome. The RFLP genotypes at one specific locus show alleles for a site on chromosomes. The linkage of desirable genes with undesirable genes is the main drawback of plant breeding.

32.4 Random amplified polymorphic DNA (RAPD)

Over the past years, researchers focused on the new exigencies for the improvement in the field of agronomy using various approaches including conventional breeding and modern molecular biotechnology. In addition, molecular marker closely associated with a trait can be used to screen and ultimately that decreases the time spent phenotyping varieties. Advances in molecular biotechnology have offered to unveil the number of DNA markers in one of which random amplified polymorphic DNA (RAPD). PCR has become the most popular and available technique to study several novel genetic analysis based on DNA amplification. RAPD is a type of PCR reaction. It uses single 10-base primer of GC-rich random sequence. They are used simultaneously for polymorphism at many sites (Grattapaglia & Sederoff, 1994). As it is PCR-based technique, amplicons using this arbitrary primer happened only after the site presents two times in a reverse direction in a span of 2000 bases. Thus RAPD polymorphism results from sequence difference in primer-binding site or target sequence which is present between priming sites. The technique starts with basic steps: (1) isolation of genomic DNA; (2) addition of single-arbitrary primer; (3) PCR; (4) gel electrophoresis of the amplified product resolved generally on 1.2%–2.0% agarose gels. The gel is stained with ethidium bromide (EtBr). Polyacrylamide gels can also be used in combination with either AgNO₃ staining, or radioactivity or fluorescently labeled primers or nucleotides; (Corley, Lim, Kalmar, & Brandhorst, 1997; Hollingsworth, Christie, Nichols, & Neilson, 1998; Huff, Peakall, & Smouse, 1993; Pammi, Schertz, Xu, Hart, & Mullet, 1994; Vejl, 1997; Weller & Reddy, 1997). (5) Determining the fragment size by comparison with the known molecular marker (Sharma, Díaz, & Blair, 2013).

The exact annealing temperature helps the random oligonucleotide primers to bind at several primer-binding sites on the complementary strand of DNA. These binding results in defected products of the priming sites. Sequence characterized amplified regions (SCARs) is based on RAPD. This is based on chromosomal reshuffles like insertions/deletions (Paran & Michelmore, 1993). Thus amplicons in a heterozygote will be identified due to the presence or absence of bands in the RAPD profile. RAPD markers are dominant in nature. In RAPD, it is not possible to distinguish between heterozygous or homozygous, as it is difficult to know that a DNA segment is amplified from which locus. There are some disadvantages of RAPD like constraints about the reproducibility of results, since this is dominant, only half of the genetic information is available and cannot detect null alleles directly.

32.4.1 Applications of random amplified polymorphic DNA

32.4.1.1 *Genetic mapping*

Short arbitrary primers are required to amplify DNA segments in RAPD. RAPD is a fast and efficient platform to perform genetic mapping with high density in many plant species such as alfalfa, fava bean, and apple (Hemmat, Weeden, Manganaris, & Lawson, 1994; Kiss, Csanadi, Kalman, Kalo, & Okresz, 1993; Torress, Weeden, & Martin, 1993).

32.4.1.2 *In development of genetic markers*

RAPD is used to study the linkage between the markers and traits. This is not required the mapping of the entire genome. RAPD technique is used to identify DNA segments to show the linkage between markers and traits (Martin, Williams, & Tanksley, 1991). It has been used in the identification of markers that are linked to disease-resistance genes in lettuce, tomato, and common bean NILs line (Adam-Blondon, Seignac, Bannerot, & Dron, 1994; Martin et al., 1991; Paran & Michelmore, 1993). This analysis is used in pooled DNA samples of NILs, which increased the gene tracking efficacy.

32.4.1.3 *In population genetics*

The RAPD technique is a simple and rapid in revealing genetic variation in DNA, thus highly used by population scientists. RAPD analysis gives more information about populations that are closely related and less information about distantly related populations. RAPD data also help in phylogenetic studies and support previously known data of RFLPs. RAPD polymorphism has been used in paternity test and kinship relationships among large populations (Smith & Williams, 1994).

32.4.1.4 *Plant breeding*

RAPD technique is used in intra-specific variation among species to screen the degree of inbreeding in commercial plants. It is used to prevent the increase of deleterious recessive allele's frequency in given populations. SCAR-transformed RAPD markers have more use in commercial plant breeding programs. RAPD markers are helpful in genetic mapping, evolutionary genetics, population genetics, and the breeding program. It is used to generate more markers in a short period.

32.5 Amplified fragment length polymorphism (AFLP)

Amplified fragment length polymorphism (AFLP) is an important technique used for fingerprinting that can apply to complex DNAs of any origin. AFLP combines both restriction digestion and PCR. First, digest the total genomic DNA and then use the digested product in the PCR reaction by ligation of adapters to digested DNA (Lynch & Walsh, 1998). AFLP can use for the selection of DNA sequences in the genome, and this phenomenon is called "genome representation." It can be created for any organism's DNA without extra cost in the development of primer/probe and sequence analysis. There is no need of high-quality DNA, low-quality DNA can also be utilized. The DNA should be devoid of any inhibitors and digestive enzymes. AFLP analysis comprises a mixture of six bases and four base cutter (EcoRI, TruI, respectively) enzymes for restriction digestion of DNA. The adapters are used to prevent the creation of restriction sites after ligation. These adapters are ligated to both fragment ends. PCR amplifies when the primers are bound to specific locations on DNA sequence. The amplicon includes adapter sequences, selective nucleotides, and complementary base pairs to the additional nucleotides. This was followed by two subsequent PCR amplifications with stringent primers. These primers are of 1–3 bases and complementary to the adapters. The first PCR is with a mixture of primer with an extra one bp, called preamplification. Further, the other amplification is carried out by 3-bp extension primer pairs. The primers are highly selective; thus primers change only by one base and are used in the AFLP extension. A primer extension up to four bases reduces the amplicon's number by factors of 4, 16, 64, and 256, respectively. AFLP fragments can be visualized by autoradiography in polyacrylamide gels (denaturing). Agarose gel, AgNO₃ staining, and next-generation sequencers can also be used to detect AFLP fragments (Fig. 32.3).

There are 50–100 amplified fragments in an AFLP fingerprint, and 80% can use as markers. Generally, AFLP needs less samples of DNA. The AFLP data show more multiplex ratio and genotyping output.

The matrix ratio (band pattern) of 1/0 is used for genotyping. The 1 is for the existence, and 0 is for the absence of a restriction fragment. This fragment was used in the detection of polymorphism by the second PCR. The characteristics of an AFLP is that a single band symbolizes two alleles at every locus. There are several bands in an AFLP gel. The band patterns could be of hundred types for every individual. It is dominance and multilocus in nature. The genotyping technology is relatively simple. The AFLP markers are also used to detect the DNA methylation. In this, pairs of restriction enzymes are used.

32.5.1 Advantages of amplified fragment length polymorphism

1. It is very productive and trustworthy (Jones et al., 1997).

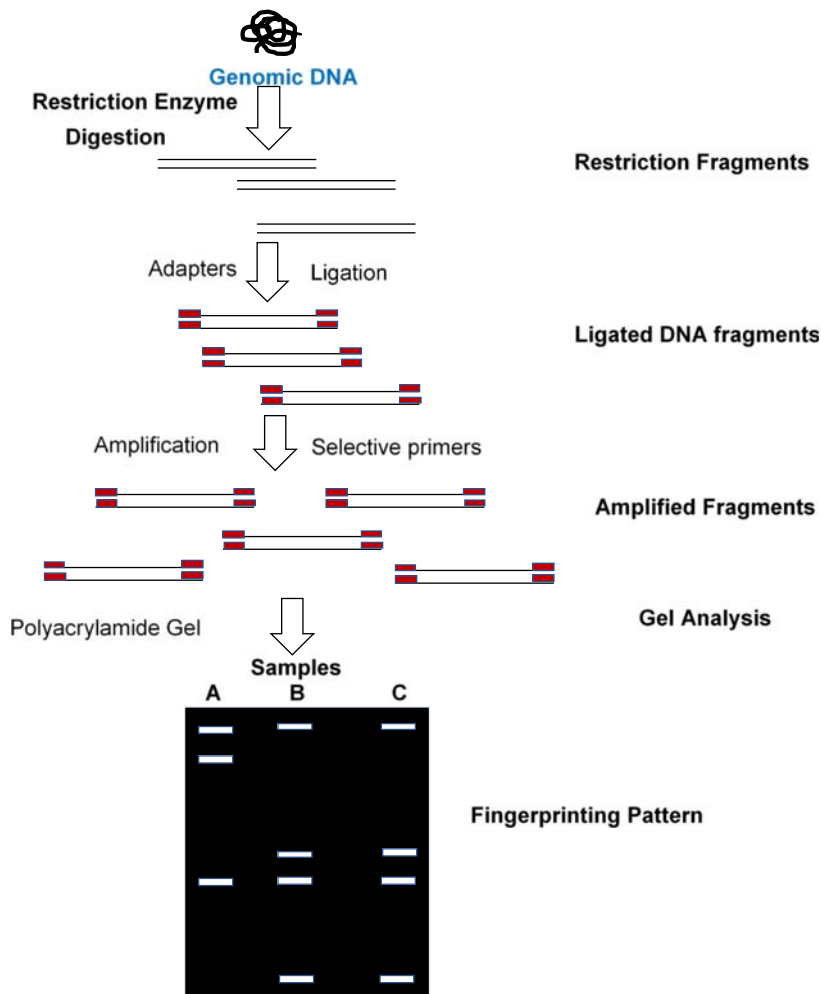


FIGURE 32.3 Schematic diagram showing steps for AFLP. AFLP, amplified fragment length polymorphism.

2. It does not require prior DNA sequence information during analysis.
3. It can use to detect a greater number of polymorphisms at different loci. It is used only one primer for the amplification as compared to RFLPs and microsatellites (Russell, Fuller, Macaulay, Hatz, & Jahoor, 1997).
4. The amplicons are comigrating and homologous for specific locus. The polyploidy species showed exceptions.
5. It can use partially degraded DNA for digestion but free of PCR inhibitors and restriction enzymes.
6. In case of high genomic heterogeneity, AFLP can use, when it is necessary to amplify many loci in outcrossing species to ascertain an accurate measure of genomic diversity.
7. It also can be used with low genetic variability. When it is necessary to amplify many loci to know the polymorphic site, AFLP is used
8. It can be used where there are no suitable established markers available.

There are many studies that showed AFLP can use in various applications (Meudt & Clarke, 2007).

32.5.2 Disadvantages of amplified fragment length polymorphism

AFLP assays have some limitations also like

1. When there are biallelic markers, it cannot be used for polymorphism.
2. Sometimes it requires quite good quality DNA for analysis.
3. AFLP markers are present in centromeres like in barley and sunflower.
4. The AFLP marker development is difficult and costly for defined locus.

32.5.3 Techniques for amplified fragment length polymorphism data analysis

32.5.3.1 Linkage mapping

AFLP data can be concomitant with different data, including RFLPs, RAPDs, and microsatellites to make linkage maps in mapping populations (such as barley, Arabidopsis, potato, and rice).

32.5.3.2 Population-based methods

There are two data analysis methods for the AFLPs. The first is based on population and used in allele frequencies comparison to divide genetic diversity. The calculation of allele frequencies from dominant markers is difficult to show the homozygous and heterozygous conditions; therefore people use the frequency of the null allele in analysis.

32.5.3.3 Phylogenetic methods

This is the second method for data analysis, individual-based, and use genetic relationships to study the individuals. AFLP data is used in phylogenetic reconstruction, closely related organisms such as ring species, recent species radiations, and crops. The combination of AFLP data and DNA sequence data showed highly robust phylogenies. This is due to the complementary effect of the different data sets which use in making the tree (Karp, 2002; Robinson & Harris, 1999).

32.5.4 Application of amplified fragment length polymorphism

The AFLP markers are applied in biodiversity studies, germplasm analysis, and construction of genetic maps. It can also use for genotyping of individuals and linkage identification, gene mapping, study of physical maps, and transcript profiling.

32.6 Simple sequence repeats (SSR)

There are various applications of molecular markers which have been discovered including correlation of genetic variations between individuals, in constructing linkage maps, population genetics and phylogenetic studies, MASs, and backcrosses. There are many molecular marker studies among many crops that have been reported in soybean, barley, and wheat (Bohn, Utz, & Melchinger, 1999; Powell et al., 1996; Russell et al., 1997). Among all molecular markers, microsatellites are important in plant genetics and breeding as it controls many genetic characters like hypervariability, relative abundance, reproducibility, multiallelic nature, codominant inheritance, chromosome-specific location, extensive genome coverage, and also high-throughput genotyping. Due to change in repeat-motifs, microsatellite markers show a high degree of allelic variation. These variations are because of replication slippage or uneven crossing-over in meiosis. Microsatellites, simple sequence repeats (SSRs) are made up of DNA sequences consisting of short, tandemly repeated nucleotide motifs. This has been found in all eukaryotic species (Tautz & Renz, 1984). Commonly plants are AT-rich repeats, whereas animals are rich in AC repeat. This seems to be the key feature that differentiates animal genomes with plant. SSRs are present in coding and noncoding region of DNA and distributed through nuclear, chloroplast, and mitochondria genome (Chung, Staub, & Chen, 2006; Kumar, Qiu, & Joshi, 2007). SSRs are showed less degree of repetition at selected locus, random distribution of genome, and high degree of length polymorphism (Zane, Bargelloni, & Patarnello, 2002). SSRs can use in high-throughput genotyping. A large number of database have been added in the public domain like for rice (<http://rgp.dna.affrc.go.jp/IRGSP>) and Arabidopsis (<http://www.arabidopsis.org>). Nowadays, expressed sequence tag (EST) databases are an important database to find candidate genes. To amplify the genic microsatellite, a locus-specific primer can be designed flanking EST- or genic SSRs. Genic SSRs are more valuable than genomic SSRs. Genic is quickly taking out from data, and they are present in genic regions of the genome. The genic SSRs are also important because they present in coding regions. In early 1993, the identification of SSRs was carried out in gene sequences of plant species by Morgante and Olivieri (Morgante & Olivieri, 1993). Microsatellites can be classified depending on repeat units, their size, and their location in the genome. Microsatellites are perfect, imperfect, and compound microsatellites. This is dependent on how the nucleotides arrange in the repeat regions. SSR's have been classified as mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotides as per number of nucleotides. The perfect repeats are tandemly duplicated repeat motif, whereas imperfect repeats have nonrepeat motifs in perfect repeats at some locations. Nuclear SSRs are the most common genomic SSRs. Microsatellites are also found in mitochondria and chloroplasts (Soranzo, Provan, & Powell, 1999; Weising & Gardner, 1999). These markers are used to know different information in the population. The change in repeat number is associated with its mutation rate. This change is dependent on time and a complex process (Pearson, Edamura, & Cleary, 2005). The change in

repeat number is caused by different processes like recombination in DNA, DNA slippage, and retro transposition. The addition and deletion of retrotransposons between the genes cause expansion of the plant genome. The slippage and recombination interaction affects SSR strength (Li, Korol, Fahima, Beiles, & Nevo, 2002) (Fig. 32.4).

32.6.1 Distribution of simple sequence repeats

SSRs is also present in the noncoding region of DNA. A large number of SSRs are present in the coding region. In cereals, only 1.5%–7.5% SSRs are located in ESTs. The dinucleotide repeats are frequent in many species. These are less in the coding region. In plants, AAG is the most frequent triplet (Li, Korol, Fahima, & Nevo, 2004). CCG is an abundant triplet in cereals. The location of SSR affects gene functions like regulating gene expression thus development of plants. If SSRs are located in a noncoding, like in 5'-untranslated regions (UTRs), the SSRs may regulate the expression of gene by affecting mRNA or protein level. In chickpea, seed weight was correlated with GA repeats variation in the 5'-UTR of the inositol mono-phosphatase gene (Dwivedi, Parida, & Chattopadhyay, 2017). The 3'-UTRs have less triplets in Arabidopsis and barley (Thiel, Michalek, Varshney, & Graner, 2003).

32.6.2 Isolation of simple sequence repeats markers

Microsatellites are present in both exons and introns. The nucleotide swap types of SSRs are abundant in noncoding regions. For the amplification of SSRs, the information about flanking DNA sequences is required to design a primer to do PCR. The amplification products are run on gel according to the size of amplicons and imagined by different dyes like EtBr, silver staining, or fluorescent dyes. The allelic variation of repeat motifs present in the microsatellite is different among genotypes

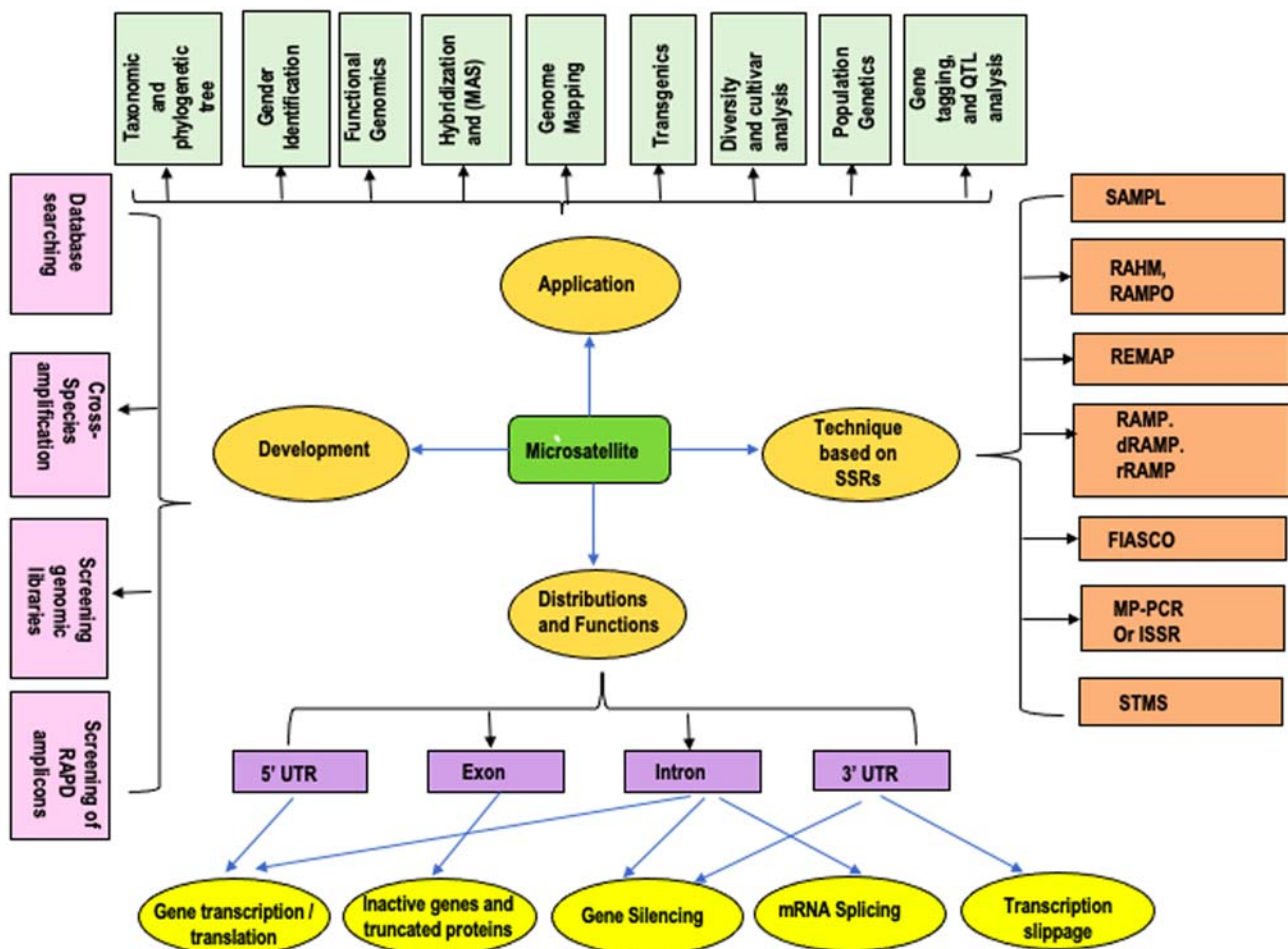


FIGURE 32.4 An overview of microsatellite markers: its development, distribution, functions, and applications in plants.

and showed polymorphisms. In potato, there are multiple SSR alleles that show heterozygosity, due to which SSR markers are more useful in potato (Milbourne, Meyer, Collins, Ramsay, & Gebhardt, 1998). There are some steps for the isolation of SSRs: (1) to create a genome library with a small insert, (2) screen the library through hybridization, (3) sequence the positive clones, (4) design of primers and PCR analysis, and (5) the last step is to identify polymorphisms in between genotypes.

32.6.3 Applications of microsatellite

32.6.3.1 Simple sequence repeats in the mapping of gene

SSRs present in the exon or intron region of the genome hardly show any information about the functions in organisms. The genic markers either have both known and unknown functions when they present in the same position of the gene and show polymorphism between different species. There are numerous EST-SSR markers and having putative functions in plants. EST-SSR markers can be useful in direct allele selection that is associated with targeted traits (Kalia, Rai, Kalia, Singh, & Dhawan, 2011). Dof homolog (DAG1) gene regulates seed germination in Arabidopsis. A homolog of this gene presented on chromosome 1B with EST-SSR markers in wheat. Two EST-SSR markers present in gene that control photo response in wheat, has been identified by Yu et al. (Yu, Dake, Singh, Benscher, & Li, 2004). The genic microsatellite markers were not present near centromere but genomic SSRs were clustered around centromere. Genic markers were found more in gene regions.

32.6.3.2 Simple sequence repeats in functional diversity

The SSRs present in the CDS of genes regulates the expression or function. However, the length of SSRs in CDS regulates the phenotypic differences as it was reported in humans for different diseases (Orr & Zoghbi, 2007). The variation present in 5'-UTR regulates the expression of the gene. Sometimes it can activate or inactivate genes or stop the proper protein formation. The SSRs present in 3'-UTR can do gene silencing or transcription slippage in many organisms. It has many roles in the examination of functional diversity between species.

32.6.3.3 Simple sequence repeats in comparative mapping

The genic SSRs markers are transferable among different species. The genomic SSRs are nontransferable between diverse species. In monocots like barley and wheat, the EST-SSRs can be used for comparative mapping (Wang, Barkley, & Jenkins, 2009). The genic SSRs are more useful as they can be used in the development of the same markers for the breeding program in many species.

32.7 Intersimple sequence repeat (ISSR)

Zietkiewicz et al. first developed inter simple sequence repeat (ISSR) technique (Zietkiewicz, Rafalski, & Labuda, 1994). The amplification of DNA segments between two microsatellite repeats regions is the main basis of ISSR. These microsatellites are at some distance and used for the amplification. The SSR markers present at the same locus have been generated for many species. Primers in this technique, called microsatellite, are of di-, tri-, tetra- or pentanucleotide. In this technique, long primers are used for amplification. There is a major drawback in the development of SSR markers because it requires the knowledge of flanking sequences. ISSR fingerprinting was established, and no prior sequence knowledge was required for analysis in this technique. Primers with repeat sequence, such as (CACT)_n, can be made with a degenerate 3'-sequence such as (CACT)₈RG or (TAGC)₆TY (R = purines, Y = pyrimidines). The resultant PCR product amplified from the sequence between two SSRs and helps to study the marker system at multiple locus useful for fingerprinting, genome mapping, and diversity analysis. These PCR products are ³²P or ³³P radiolabeled with the help of end-labeling or PCR incorporation. The PCR product is separated on a polyacrylamide sequencing gel and further used in autoradiographic visualization. A typical reaction yields 20–100 bands per lane depending on the species and primer. They are segregating by simple Mendelian laws of inheritance, thus characterized as dominant markers (Fang & Roose, 1997; Tsumura, Ohba, & Strauss, 1996). These can also be used as codominant markers. ISSRs are simple, easy markers as compared to RAPD (Chatterjee, Vijayan, & Roy, 2004; Kar, Vijayan, & Mohandas, 2005). Moreover, they are dominant markers in nature. They are less reproducible and show homology with comigrating amplification products (Semagn, Bjornstad, & Ndjioudjop, 2006).

32.7.1 Advantages of intersimple sequence repeat markers

1. The major advantage of ISSR is no previous requisite of the DNA sequence for analysis.
2. This is simpler and more reliable than other techniques, as the PCR products are specific.

3. This is fast as it can do simultaneous assessment of various loci.
4. It can distinguish closely related species.
5. It is less costly and time-consuming, as it is devoid of cloning and characterization.
6. It is easy to use, as less steps to be required as compared to AFLP and cost-effective.

32.7.2 Disadvantages of intersimple sequence repeat markers

ISSR assays have some limitations also like

1. It is dominant in nature.
2. It leads to abstruse fingerprints, as sometimes the ISSR primers have not much specificity to the genome.
3. The quality of DNA is important; the poor-quality DNA gives poor result.

32.7.3 Application of intersimple sequence repeat markers

Scientists have used fingerprinting based on this technique in different plant species. Previous scientists have demonstrated that there is higher level of polymorphism when ISSR markers used when compared with RFLP or RAPD analyses. It has been also used for genetic diversity analysis. It is used to study genetic diversity analysis in sorghum and maize and cultivar identification in chrysanthemum (Chatterjee et al., 2004). It is used in the forest species *Cryptomeria japonica* and *Pseudotsuga menzeisii* (Reddy, Sarla, Neeraja, & Siddiq, 2000; Tsumura et al., 1996).

32.8 Single-nucleotide polymorphism (SNP)

If one base pair variation is detected in the DNA sequence of different species, then it is called SNPs. This is the most abundant marker in both animal and plant. The advanced sequencing technology like NGS can produce huge sequencing data, which make the easy way to identify SNPs. Nowadays, numerous sequencing technologies are available to develop huge genotyping data, which conceded easy, efficient fast identification of SNP markers in many plants (Ganal, Wieseke, Luerssen, Durstewitz, & Graner, 2014). High-density SNPs markers are usually used to analyze population structure and genetic diversity. These are used to develop genetic maps and study the genome-wide association study (GWAS). In the same species, allelic variations within a genome can be divided into three categories (1) SSRs, differences in repeat number, (2) InDels, insertions/deletions, and (3) SNPs. Due to their biallelic nature, SNPs show less polymorphism than SSRs. This drawback can easily compensate by next-generation sequencing (NGS), high-throughput automation. In NGS, parallelize DNA sequencing can be done and this helps to read molecules of thousands of genetic materials simultaneously. Nowadays, there are several methods that have low genome sequence and had been successfully developed. These are reduced representation libraries (Hyten et al., 2010), GBS (genotyping-by-sequencing) (Elshire, Glaubitz, Sun, Poland, & Kawamoto, 2011), RAD (restriction-site associated) sequencing (Bus, Hecht, Huettel, Reinhardt, & Stich, 2012), and SLAF-seq (single-locus amplified fragment sequencing) (Zhang et al., 2013). SLAF-seq is an easy, efficient, accurate, and cost-effective method for the development of SNPs and InDels. SNPs discoveries in polyploid crops are difficult like in coaon, canola, and wheat. SNP discovery in allopolyploids plant confides upon differentiation in sequence variation (Thomson, 2014). The haplotype information and allelic frequency are together used to differentiate between homologous SNPs and homoeologous loci. SNP genotyping can be used as genomic tools that developed new approaches in mapping complex traits.

32.8.1 Single-nucleotide polymorphism detection

SNP can be detected by two methods: (1) *in silico* and (2) *in vitro* techniques. *In silico* methods are easy for SNPs mining in species with the known genome sequences. There are many software and databases for SNPs mining in plants. Illumina GA/Solexa, SOLiDTM, Oxford and Nanopore are advanced sequencing platform to generate a large number of SNPs. Reference and *de novo* are two type of sequence data. There are three steps to identify SNPs from sequencing (1) group the sequence reads, (2) read alignments, (3) sequence variants scanning. There are several softwares or databases for SNP mining in plants such as dbSNP, POLYMORPH, SNiPlay, and IRIS (Table 32.3).

32.8.2 *In vitro* techniques

This can be divided into three categories (1) nonsequencing technique-like cleaved amplified polymorphic sequences, single-strand conformation polymorphism, temperature-gradient gel electrophoresis, and denaturing-gradient gel

electrophoresis. (2) Sequencing-reduced representation shotgun, bacterial artificial chromosome, and PAC (P1-derived artificial chromosome). (3) Re-sequencing MALDI-TOF/MS and sequencing.

32.8.3 Single-nucleotide polymorphism application

The genomic variation is much important for plant breeding and genetics. The SNPs cover large populations to identify desired lines for different traits. Today, SNPs is widely used in plant breeding, feature mapping, and cost-effective analysis of germplasm. The use of SNPs improves the understanding of plant genetics thus changing the strategy for new varieties. SNPs can use in the detection of correlations between genotypes and phenotypes. It is used to identify common diseases with complex genetics, genetic diagnostics, genetic diversity analysis, construction of genetic map, association mapping (AM) by linkage disequilibrium (LD), phylogenetic analysis, and cultivar identification (Brachi, Morris, & Borevitz, 2011). To identify new alleles, genetic diversity information can be used in plant breeding. SNPs can be used in plant phylogenetic and evolutionary research (Fournier-Level, Lacombe, Le Cunff, Boursiquot, & This, 2010). SNPs present in nuclear and chloroplast gene regions are a rich source of phylogeny in plants. SNP is good to study the genetic diversity in domesticated populations. SNPs can cause phenotypic diversity like plants/flowers/fruits color, ripening timing, time to first flower, fruit size, grain yield, crop quality, or various abiotic and biotic stress tolerance (Huq, Akter, Nou, & Kim, 2016). SNPs present in the exon of a gene and can changes the amino acids. SNP can also silence the gene. There are many advantages of SNP markers like highly flexible and fast as they provide huge data for analysis. The high-quality reference genome gives the entire SNP catalog for each species. SNPs are also used to detect the variation in UTRs.

32.8.4 Diversity array technology (DArT Seq)

This technique is highly reproducible and based on microarray hybridization. This is a technique, which helps to identify polymorphic loci (hundreds to thousands) in the genome. For the detection by this method, it is not needed the previous sequence information (Wenzl, Carling, & Kudrna, 2004). It is very economical and highly throughput. One reaction can identify several loci by this technology. This requires little amount of DNA (55–105 ng genomic DNA). An identical platform is utilized for the scoring and discovery of markers in this technique. This is not required specific assays for genotyping. It is just need the assembly of polymorphic markers into one genotype array. These markers are used for genotypic reaction (Huttner, Wenzl, & Akbari, 2005).

32.9 Quantitative trait loci (QTL)

Many important traits like yield traits, quality, root architecture, and disease resistance are regulated by multiple genes and therefore are known as quantitative traits. The QTL identification is also based on DNA (or molecular) marker besides conventional phenotypic evaluation. DNA markers have been used in the construction of linkage maps in agricultural research. Linkage maps is used in QTL analysis to identify the chromosomal regions having genes (Mohan, Nair, & Bhagwat, 1997). QTL mapping includes linkage maps construction followed by analysis of QTL (McCouch & Doerge, 1995; Paterson & Landes Company, 1996; Zeng, 1994). There are many steps in QTL mapping (Fig. 32.5).

32.9.1 Molecular markers

DNA marker can be divided into three categories (1) hybridization-based; (2) PCR-based, and (3) DNA sequence-based (Gupta, Varshney, Sharma, & Ramesh, 1999) (Table 32.1).

DNA markers are quite useful if they show differences between the same or different species. These markers are different between species and are known as polymorphic markers, whereas monomorphic markers are same between genotypes. Polymorphic markers are codominant or dominant in nature. The markers can discriminate between homozygous and heterozygous plant.

32.9.2 Construction of genetic linkage maps

A linkage map, a “road map” on the chromosomes is derived from different individuals. The linkage maps are used to construct QTL (or “genetic”) map. In QTL mapping, genes and markers segregate via chromosome recombination (called crossing-over) (Paterson & Landes Company, 1996). Linkage map constructions are involved three things (1) mapping population; (2) polymorphism, and (3) linkage analysis.

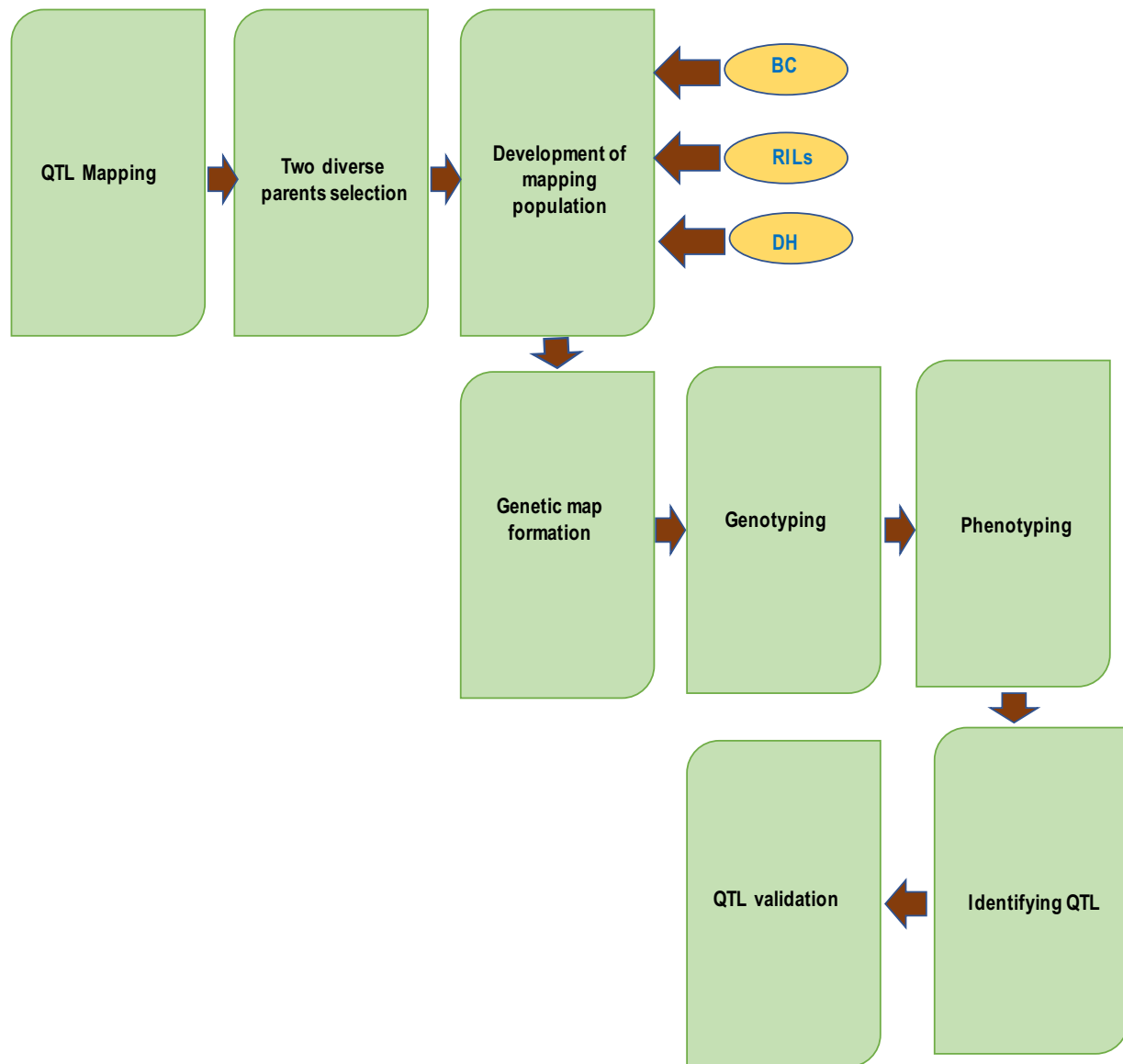


FIGURE 32.5 Steps involved in QTL mapping. *QTL*, quantitative trait loci.

32.9.3 Mapping population

A segregating plant population that is derived from sexual reproduction is required for the linkage map construction. The mapping population should be made with contrasting parents. Population sizes of a mapping population may range from 100 to 300 individuals (Zeng, 1994).

32.9.4 Identification of polymorphism

Identifying DNA markers is the second step, which shows differences between parents in the construction of a linkage map. The polymorphic markers will be checked in the whole population, which includes their parents. This process is called genotyping.

32.9.5 Linkage analysis of markers

The last step of the construction of linkage map is linkage analysis. This involves coding data for DNA marker on each mapping population individual and analyzing linkage through software. Linkage is generally calculated using odds ratios and called a logarithm of odds (LOD) value or LOD score (Risch, 1992).

TABLE 32.1 Types of molecular/DNA markers

DNA marker	Single/ multiple loci	Degree of dominance	Polymorphism source (mutation)	Polymorphism level	Abundance	Laboratory technique
RFLPs	Single	Codominant	Point mutation	Lower	Moderate	Southern blot; agarose gel
RAPDs	Multiple	Dominant	Point mutation	Lower	Little	PCR; agarose gel
AFLPs	Multiple	Dominant	Point mutation	Lower	Moderate	PCR; acrylamide gel
SSRs	Single	Codominant	Variation in the number of the repeats	Higher	Moderate	PCR; acrylamide gel
SNPs	Single	Codominant	Point mutation (with sequence information)	Lower to Higher	Very high	Primer extension; chips
ISSR	Single	Dominant	Variation in the number of the repeats	Higher	Moderate	PCR; agarose gel
DART	Single	Dominant	Point mutation	Higher	Very high	Microarray
Retrotransposons	Single	Dominant	Point mutation	Higher	High	PCR; agarose gel

32.9.6 Genetic distance and mapping functions

The chance of recombination is based on the distance between two markers. Distance is calculated by the recombination frequency between genetic markers. The recombination frequency is presented by centiMorgans (cM) (Hartl & Jones, 2005).

32.9.7 Quantitative trait loci analysis

The phenotype and genotype association of markers is the key component in QTL analysis. The genotypic groups are based on the presence or absence of a particular DNA marker locus (Young, 1996). If marker and QTL are closely linked, then the chance of recombination is less. Therefore the marker and QTL will be generally inherited together in the next generation.

32.9.8 Quantitative trait loci detection

There are commonly three ways to detect QTLs: (1) single-marker, (2) simple interval, and (3) composite interval (Liu & Wu, 1998; Tanksley, 1993). The is single-marker analysis in which one can use single marker for QTL. This method uses different statistical methods like *t*-tests and linear regression. Q Gene and MapManager software are mostly used for this analysis (Manly, Cudmore Robert, & Meer, 2001).

On the other hand, the simple interval mapping method first make linkage maps and then simultaneously check linkage intervals between markers along all chromosomes (Lander & Botstein, 1989). But nowadays composite interval mapping (CIM) is easy and more in use for QTL mapping. It includes genetic markers in the statistical model and interval mapping with linear regression. QTL Cartographer, MapManager QTX, and PLABQTL are used to study CIM (Basten, Weir, & Zeng, 2001; Manly et al., 2001; Utz & Melchinger, 1996).

32.9.9 Advantages and disadvantages of quantitative trait loci mapping

It is utilized to spot the genes of interest that control the particular trait in plants. It is very beneficial for the genome-wide detection of QTLs in plants. The QTL mapping can detect the genes controlling disease resistance in plant because diseases are a big apprehension of agriculture. Besides advantages, there are some disadvantages of QTL mapping like

lower number of recombination events, less allelic diversity (Price, 2006). It is also a time-consuming when we developed a mapping population. It is less specific in the detection of QTLs for a given population.

32.10 Association mapping

In AM, molecular markers are associated with a phenotypic trait. AM shows a correlation between the polymorphic marker and the trait of interest (Jannink & Walsh, 2002; Zhang, Zhong, & Shahid, 2016). In comparison with linkage mapping, it is time saving technique, provides greater mapping resolution, and shows more recombination events. AM expedites the documentation of a higher number of alleles. AM is based on LD.

32.10.1 Linkage disequilibrium

LD is when alleles are associated nonrandomly at different loci. LD showed the increased or decreased haplotypes frequency in a population. In LD, gametic disequilibrium or gametic phase disequilibrium can be denoted as $P_{AB} \neq P_A \times P_B$, where P_{AB} is a frequency of haplotypes of alleles AB; P_A is a frequency of haplotypes A, and P_B is a frequency of haplotypes B. GOLD, TASSEL, and R are the most commonly used software for LD pattern (Bradbury, Zhang, & Kroon, 2007). Mutation and recombination are important for the significant LD.

32.10.2 Methods of association mapping

A wide range of genetically diverse populations are used in AM. First phenotyping has been done for the population at a different time point, different locations, and in different environments. After phenotyping, genotyping with favorable markers is done. Further, the populations' structure and kinship matrix are calculated. Further, the phenotyping and genotyping data are correlated with help of different software. TASSEL is commonly used software for AM. The detailed method is as follows Fig. 32.6.

32.10.3 Class of association mapping

AM can divide into two classes: (1) candidate-gene-based and (2) genome-wide association.

32.10.3.1 Candidate-gene-based

This technique is used to study the correlation between the DNA polymorphism present in a gene and an interested trait (Sehgal, Singh, & Rajpal, 2016). In this technique, biologically relevant candidates are selected based on their evolutionary

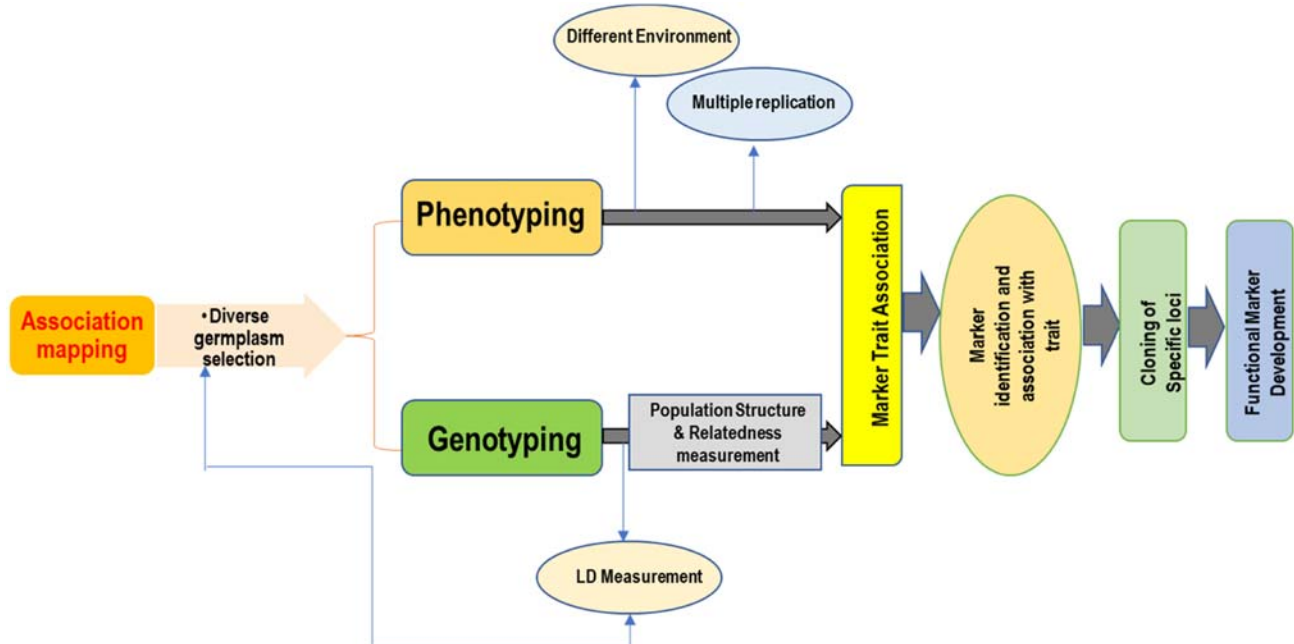


FIGURE 32.6 Methods to study the association mapping.

data. This technique uses SNPs present within specific genes and between lines. The SNPs present in exon, promoter, and 5'/3'-UTRs are the important factors to investigate the candidate gene. This technique has been widely used for the identification of QTLs in the last several years. This is also used for the development of functional markers (Lau, Rafii, & Ismail, 2015).

32.10.3.2 Genome-wide association study

This technique is used in various species. It uses the GBS. This is used to study the genetic variations in plants. Normally, recombinant inbred lines are used for GWAS. For QTLs, a large population is required to obtain a high-resolution genetic map. This technique is used to investigate the small haplotype blocks that are correlated with quantitative traits. This is a cheap and high-throughput method and used in different crops like millet, maize, wheat, rice, sorghum, and chickpea (Jia & Zhao, 2014).

32.10.4 Association mapping in the breeding program

AM used superior alleles for breeding practices for introgressive hybridization into elite germplasm from different individuals. There are many examples of the use of AM in breeding programs. Most studied characters are abiotic stress and yield-related characters. RHM (regional heritability mapping) and GWAS for productivity, lodging, and plant architecture have been carried out by Resende et al. (2018). They were used 188 germplasms of common bean. This includes three markers used for trait study with help of GWAS, and 145 markers were identified using RHM along chromosomes 5. Liu, Bayer, Druka, and Russell (2014) have identified a total 122 and 134 QTL for different traits of cotton in two environments using GWAS. Patishtan, Hartley, Fonseca de Carvalho, and Maathuis (2018) performed GWAS in a panel of 306 rice germplasm to identified transcription factors and components of the ubiquitination pathway (Patishtan, Hartley, Fonseca de Carvalho, & Maathuis, 2018). PIP2, RD2, and PP2C genes were found to be significant for abiotic stress resistance in cotton (Hou et al., 2018). Zhang and Yuan (2019) showed AM in maize (300 inbred lines) from different habitats.

32.11 Marker-assisted selection (MAS)

MAS includes molecular marker and traditional breeding program. In MAS, a single cross is made. The steps involved in MAS is shown in Fig. 32.7. There are many steps in MAS as following:

- Select the parents for crossing that have DNA marker alleles for the trait of interest.
- The second plant F1 population is detected for the marker alleles thus eliminating false hybrids.
- Screen F2 population and store data for the individuals having the desired alleles.
- Select F3 individuals for desired marker alleles and traits.
- Screen F4 and F5 generation for marker to find homozygous lines and evaluate the best lines for phenotypic trait of interest
- Evaluate the selected lines for characters of interest like yield, quality, and resistance.

In MAS, theoretically all the QTLs could be taken for analysis that contributes to the trait of interest. The MAS efficiency depends on QTL number, as the number increases the efficiency decreases thus their heritability also decreases (Moreau et al., 1998). The number of genes/QTLs are also controlled the efficiency of MAS. Generally, more than three QTLs are not good for MAS. There is some report in which scientists used more than three QTLs. Five QTLs were used through marker-assisted introgression for the improvement of fruit quality traits in tomato (Lecomte, Duffé, Buret, Servin, & Hospital, 2004). Two markers for a single QTL are good from both an effectiveness and efficiency point of view. The efficiency of MAS depends on recombination frequency inversely.

32.11.1 Application of marker-assisted selection

In crop plants, many traits are governed by QTLs like diseases/pests resistance, self-incompatibility, and male sterility. MAS has been used in soybean cyst nematode (*Heterodera glycine* Inchinoe) for major genes. As some traits are controlled by QTLs so strong, QTL–environmental interaction control the phenotypic expression. A saturated linkage map showed both targeted and linked QTLs. MAS can be used to improve the plants.

MAS can be used to make disease-resistance plants like in wheat; markers for MAS and powdery mildew resistance genes like Pm4a, Pm5e, and other Pm genes have been identified (Huang et al., 1997; Ma, van der Does, & Borkovich, 2010).

MAS is also used for the mapping of desired traits involved in tolerance to drought stress like osmotic adjustment, root penetration and morphology, carbon isotope discrimination, and phenological traits like anthesis-silking interval in maize.

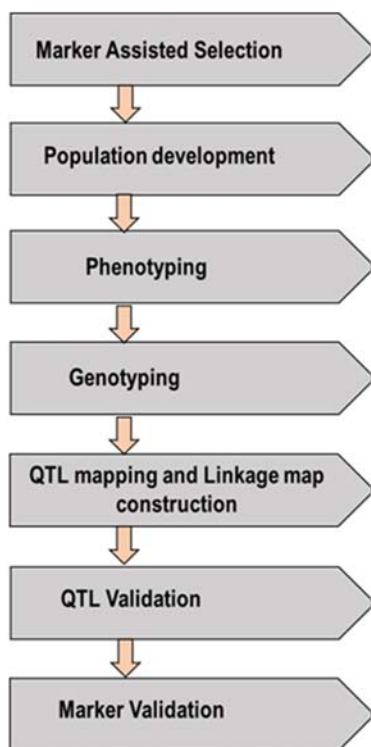


FIGURE 32.7 Key steps involved in MAS study. MAS, marker-assisted selection.

TABLE 32.2 Comparison of key features between different sequencing platforms.

Topographies	Roche 454	Ion torrent	Illumina
Sequence method	Pyrosequencing	Semiconductor	Synthesis
PCR approach	EmPCR	EmPCR	BridgePCR
Sequencing-paired end	No	No	Yes
Read count (bp)	350–1000	~200	100–250

32.12 Bioinformatics intervention in molecular markers

NGS has significantly improved plant genomics. NGS technology is used to discover, validate, and assess the molecular markers on a large scale. NGS technology has produced a huge amount of data that can be used for the mining of markers in plant.

These markers further can be used for genetic mapping, studies of association, analysis of diversity among populations, and MAS. Advanced bioinformatics tools and databases are required to mine the data that result in the discovery of molecular markers. NGS technology helped in the discovery of markers and genotyping of these markers at a very high density. These markers are used for complete GWAS. Different NGS technologies are commonly used for marker discovery like Roche/454 sequencing, Ion torrent: Proton/PGM sequencing, Illumina (Solexa) sequencing, SOLiD sequencing, and Pacific Biosciences (Table 32.2). From recent studies, it has been found that shotgun sequencing of a genome or transcriptome through NGS platform is the easy way to mine SNP or SSR marker. Genome assembly is the first step to find the markers in NGS platform. There are many softwares for genome assembly like CLC Genomics Workbench, Velvet, and SeqManNGen (DNASTAR) (Table 32.3).

32.13 Software for simple sequence repeats discovery

There are many softwares to screen the SSR in the entire genome. There are some tools for SSR like MicroSatellite; mreps, the windows-based SSR locator, WebSat, and Msatfinder 2.0. Search for Tandem Approximate Repeats

TABLE 32.3 Different software: software uses for assembly of sequences, in the identification of SSRs and SNPs.

	Assembly software	
Name	Technology	Website/References
CLC Genomics Workbench	Sanger, 454, Illumina, Ion torrent	http://www.clcbio.com/
Velvet	Sanger, 454, Illumina	http://www.ebi.ac.uk/~zerbino/
AbySS	Sanger, 454, Illumina, Ion torrent	http://www.bcgsc.ca/platform/bioinfo/software/abyss
SeqManNgen	Sanger, 454, Illumina, Ion torrent	http://www.dnastar.com/t-
MIRA	Sanger, 454, Illumina, Ion torrent	http://sourceforge.net/apps/mediawiki/mira-assembler/
TMAP	Ion torrent	http://www.ioncommunity.lifetechnologies.com/
NextGENe	Sanger, 454, Illumina, Ion torrent	http://softgenetics.com/NextGENe.html
TopHat	454, Illumina	Lorenc, Boskovic, Stiller, Duran, and Edwards (2012)
	SSR tools	
SSRPrimerII		http://www.appliedbioinformatics.com
MicroSATellite (MISA)		http://pgrc.ipk-gatersleben.de/misa/
SSR identification tool (SSRIT)		Kantety, La Rota, Matthews, and Sorrells (2002)
Tandem repeat occurrence locator (TROLL)		Castelo, Martins, and Gao (2002)
SSRSEARCH		ftp://ftp.gramene.org/pub/gramene/
RepeatMasker		http://www.mendeley.com/
Msatfinder		http://www.genomics.ceh.ac.uk/
RepeatMasker		http://www.mendeley.com/
	SNP tools	
SOAP2		http://soap.genomics.org.cn/
Samtools		http://samtools.sourceforge.net/
GATK		http://www.broadinstitute.org
MaCH		http://genome.sph.umich.edu/
Qcall		ftp://ftp.sanger.ac.uk/pub/
IMPUTE2		http://mathgen.stats.ox.ac.uk/
GigaBayes		http://bioinformatics.bc.edu/
SNPdetector		Zhang et al. (2005)
Geneious		http://www.geneious.com
SGSautoSNP		Lorenc et al. (2012)
QualitySNP		Tang, Vosman, Voorrips, van der Linden, & Leunissen (2006)
PolyScan		Chen, McLellan, Ding, Wendl, and Kasai (2007)

(STAR), a mining tool, is used for the identification of repeats motif (Ruperao & Edwards, 2014). Some tools like msatfinder, E-TRA, msatcommander, and MISA are used for SSR finding from NGS data (Table 32.3).

32.14 Software for single-nucleotide polymorphism discovery

Because of their abundance, SNPs have emerged as the markers of choice and are used in various breeding programs. SNPs have huge potential in crop improvement programs. There are various methods for the detection and genotyping of SNPs.

There are many SNP discovery software programs such as Consensus Assessment of Sequence And Variation (CASAVA), NextGENe (<http://www.softgenetics.com/>), CLC Genomics Workbench, Biomatters, Geneious, SNPdector, and ACCUSA (Ruperao & Edwards, 2014).

The SNP discovery software AutoSNPdb can be used for both Sanger and Roche 454 (84). MAQ, using the alignment quality, used to predict SNPs Table 32.3.

The advancement of bioinformatics tool will be necessary for the analysis of genomic data obtained from sequencing to mine the markers. NGS technology has changed the way of study of markers for genotyping. It helps to generate markers for comprehensive association studies in genome. Through the combination of NGS technology and bioinformatics, numerous biological questions can be studied like recombination breakpoints for trait association and used in genomic selection for crop improvement.

References

- Adam-Blondon, A. F., Sevignac, M., Bannerot, H., & Dron, M. (1994). SCAR, RAPD and RFLP markers linked to a dominant gene (Are) conferring resistance to anthracnose in common bean. *Theoretical and Applied Genetics*, 88, 865–870.
- Basten, C., Weir, B., & Zeng, Z. B. (2001). *QTL cartographer*. Raleigh, NC: Department of Statistics, North Carolina State University.
- Bohn, M., Utz, H. F., & Melchinger, A. E. (1999). Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. *Crop Science*, 39, 228–237.
- Brachi, B., Morris, G., & Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, 23(2), 10.1186.
- Bradbury, P. J., Zhang, Z., Kroon, D. E., et al. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633–2635.
- Bus, A., Hecht, J., Huettel, B., Reinhardt, R., & Stich, B. (2012). High throughput polymorphism detection and genotyping in Brassica napus using next-generation RAD sequencing. *BMC Genomics*, 13, 281.
- Cadalen, T., Sourdille, P., Charmet, G., Tixier, M. H., Gay, G., Boeuf, C., Bernard, S., Leroy, P., & Bernard, M. (1998). Molecular markers linked to genes affecting plant height in wheat using a doubled haploid population. *Theoretical and Applied Genetics*, 96, 933–940.
- Castelo, A. T., Martins, W., & Gao, G. R. (2002). TROLL—tandem repeat occurrence locator. *Bioinformatics*, 18, 634–636.
- Chatterjee, S. N., Vijayan, K., Roy, G. C., et al. (2004). ISSR profiling of genetic variability in the ecotypes of *Antheraea mylitta* Drury, the tropical tasar silkworm. *Russian Journal of Genetics*, 40(2), 152–159.
- Chen, K., McLellan, M. D., Ding, L., Wendl, M. C., Kasai, Y., et al. (2007). PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Research*, 17, 659–666.
- Chung, A. M., Staub, J. E., & Chen, J. F. (2006). Molecular phylogeny of Cucumis species as revealed by consensus chloroplast SSR marker length and sequence variation. *Genome*, 49, 219–229.
- Collard, B. C., & Mackill, D. J. (2008). Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B*, 363(1491), 557–572.
- Corley, S., Lim, G. E., Kalmar, C. J., & Brandhorst, B. P. (1997). Efficient detection of DNA polymorphisms by fluorescent RAPD analysis. *BioTechniques*, 22, 690–692.
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., et al. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7), 499–510.
- Duzyaman, E. (2005). Phenotypic diversity within a collection of distinct okra (*Abelmoschus esculentus*) cultivars derived from Turkish landraces. *Genetic Resources and Crop Evolution*, 52, 1019–1030.
- Dwivedi, V., Parida, S. K., & Chattopadhyay, D. (2017). A repeat length variation in myo-inositol monophosphatase gene contributes to seed size trait in chickpea. *Scientific Reports*, 7, 4764.
- Eagles, H. A., Bariana, H. S., Ogbonnaya, F. C., et al. (2001). Implementation of markers in Australian wheat breeding. *Crop Pasture Sci*, 52(12), 1349–1356.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., & Kawamoto, K. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, 6, e19379.
- Fang, D. Q., & Roose, M. L. (1997). Identification of closely related citrus cultivars with inter-simple sequence repeat markers. *Theoretical and Applied Genetics*, 95(3), 408–417.

- Fournier-Level, A., Lacombe, T., Le Cunff, L., Boursiquot, J. M., & This, P. (2010). Evolution of the VvMYbA gene family, the major determinant of berry colour in cultivated grapevine (*Vitis vinifera* L.). *Heredity*, *104*, 351–362.
- Ganal, M. W., Wieseke, R., Luerksen, H., Durstewitz, G., Graner, E. M., et al. (2014). High-throughput SNP profile of genetic resources in crop plants using genotyping arrays. In R. Tuberosa, A. Graner, & E. Frison (Eds.), *Genomics of plant genetic resources* (pp. 113–130). Dordrecht: Springer.
- Grattapaglia, D., & Sederoff, R. (1994). Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics*, *137*, 1121–1137.
- Gupta, P. K., Varshney, R. K., Sharma, P. C., & Ramesh, B. (1999). Molecular markers and their applications in wheat breeding. *Plant Breeding*, *118*, 369–390.
- Hartl, D. L., & Jones, E. W. (2005). *DNA structure and DNA manipulation. Genetics: analysis of genes and genomes* (5th ed.), pp. 36–85. Sudbury: Jones and Bartlett Pub, Ch. 2.
- Hemmat, M., Weeden, N. F., Manganaris, A. G., & Lawson, D. M. (1994). Molecular marker linkage map for apple. *Journal of Heredity*, *85*, 4–11.
- Hoisington, D., Bohorova, N., Fennell, S., Khairallah, M., Pellegrineschi, A., & Ribaut, J. M. (2002). In B. C. Curtics, S. Rajaram, & H. Gomez (Eds.), *The application of biotechnology in wheat improvement and production*. Rome: FAO.
- Hollingsworth, W. O., Christie, C. B., Nichols, M. A., & Neilson, H. F. (1998). Detect ion of variation among and within asparagus hybrids using random amplified DNA (RAPD) markers. *New Zealand Journal of Crop and Horticultural Science*, *26*, 1–9.
- Hou, S., et al. (2018). Genome-wide association studies reveal genetic variation and candidate genes of drought stress-related traits in cotton (*Gossypium hirsutum* L.). *Frontiers in Plant Science*, *9*.
- Huang, N., Angeles, E. R., Domingo, J., Magpantay, G., Singh, S., Zhang, G., et al. (1997). Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. *Theoretical and Applied Genetics*, *95*, 313–320.
- Huff, D. R., Peakall, R., & Smouse, P. E. (1993). RAPD variation within and among natural populations of outcrossing buffalograss [*Buchloe dactyloides* (Nutt.) Engelm]. *Theoretical and Applied Genetics*, *86*, 927–934.
- Huq, M. A., Akter, S., Nou, I. S., Kim, H. T., et al. (2016). Identification of functional SNPs in genes and their effects on plant phenotypes. *Journal of Plant Biotechnology*, *43*, 1–11.
- Huttner, E., Wenzl, P., Akbari, M., et al. (2005). Diversity arrays technology: A novel tool for harnessing the genetic potential of orphan crops. In I. Serageldin, & G. J. Persley (Eds.), *Discovery to delivery: BioVision Alexandria; Proceedings of the Conference of the World Biological Forum; Alexandria, Egypt* (pp. 145–155). Wallingford: CABI.
- Hyten, D. L., Cannon, S. B., Song, Q., Weeks, N., Fickus, E. W., Shoemaker, R. C., et al. (2010). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics*, *11*, 38.
- Jannink, J. L., & Walsh, B. (2002). Association mapping in plant populations. In M. S. Kang (Ed.), *Quantitative genetics, genomics and plant breeding* (pp. 59–68). Oxford: CAB International.
- Jia, P., & Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Human Genetics*, *133*(2), 125–138.
- Jiang, G. L. (2013). Molecular markers and marker-assisted breeding in plants. In S. B. Andersen (Ed.), *Plant breeding from laboratories to fields* (pp. 45–83). Rijeka: InTech.
- Jonah, P. M., Bello, L. L., Lucky, O., A. Midau, A., et al. (2011). Review: The importance of molecular markers in plant breeding programmes. *Global Journal of Science Frontier Research*, *11*, 5.
- Jones, C. J., et al. (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Molecular Breeding*, *3*, 381–390.
- Joshi, M., & Deshpande, J. D. (2011). Polymerase chain reaction: methods, principles and application. *International Journal of Biomedical Research*, *2*(1), 81–97.
- Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., & Dhawan, A. K. (2011). Microsatellite markers: An overview of the recent progress in plants. *Euphytica*, *177*, 309–334.
- Kantety, R. V., La Rota, M., Matthews, D. E., & Sorrells, M. E. (2002). Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Molecular Biology*, *48*, 501–510.
- Kar, P. K., Vijayan, K., Mohandas, T. P., et al. (2005). Genetic variability and genetic structure of wild and semi-domestic populations of tasar silkworm (*Antheraea mylitta*) ecorace Daba as revealed through ISSR markers. *Genetica*, *125*(2–3), 173–183.
- Karaköy, T., Baloch, F. S., Toklu, F., et al. (2014). Variation for selected morphological and quality-related traits among 178 faba bean landraces collected from Turkey. *Plant Genetic Resources*, *12*(01), 5–13.
- Karp, A. (2002). The new genetic era: will it help us in managing genetic diversity? In *Managing Plant Genetic Diversity* (J. M. M. Engels, V. Ramanatha Rao, A. H. D. Brown, and M. T. Jackson, eds.); pp. 43–56.
- Kebriyae, D., Kordrostami, M., Rezadoost, M. H., et al. (2012). QTL analysis of agronomic traits in rice using SSR and AFLP markers. *Notulae Scientia Biologicae*, *4*(2), 116–123.
- Kiss, G. B., Csanadi, G., Kalman, K., Kalo, P., & Okresz, L. (1993). Construction of a basic linkage map for alfalfa using RFLP, RAPD, isozyme and morphological markers. *Molecular Genetics and Genomics*, *238*, 129–137.
- Kumar, R., Qiu, J., Joshi, T., et al. (2007). Single feature polymorphism discovery in rice. *PLoS One*, *2*, e284.
- Lander, E. S., & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, *121*, 185–199.
- Lau, W. C., Rafii, M. Y., Ismail, M. R., et al. (2015). Review of functional markers for improving cooking, eating, and the nutritional qualities of rice. *Frontiers in Plant Science*, *6*, 832.

- Lecomte, L., Duffé, P., Buret, M., Servin, B., Hospital, F., et al. (2004). Marker-assisted introgression of five QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. *Theoretical and Applied Genetics*, *109*, 568–668.
- Li, Y. C., Korol, A. B., Fahima, T., & Nevo, E. (2004). Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution*, *21*, 991–1007.
- Li, Y. C., Korol, A. B., Fahima, T., Beiles, A., & Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, *2002*(11), 2453–2465.
- Liu, H., Bayer, M., Druka, A., Russell, J. R., et al. (2014). An evaluation of genotyping by sequencing (GBS) to map the *Breviaristatum-e* (ari-e) locus in cultivated barley. *BMC Genomics*, *15*, 1.
- Liu, X. C., & Wu, J. L. (1998). SSR heterotic patterns of parents for making and predicting heterosis. *Molecular Breeding*, *4*, 263–268.
- Lorenc, M. T., Boskovic, Z., Stiller, J., Duran, C., & Edwards, D. (2012). Role of bioinformatics as a tool for oilseed Brassica species. In D. Edwards, I. A. P. Parkin, & J. Batley (Eds.), *Genetics. Genomics and breeding of oilseed Brassicas* (pp. 194–205). New Hampshire: Science Publishers Inc.
- Lorenc, M. T., Hayashi, S., Stiller, J., Lee, H., Manoli, S., Ruperao, P., et al. (2012). Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology*, *1*, 370–382.
- Lynch, M., & Walsh, J. B. (1998). *Genetics and analysis of quantitative traits*. Sunderland: Sinauer Associates.
- Ma, L. J., van der Does, H. C., Borkovich, K. A., et al. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*, *464*, 367–373.
- Manly, K. F., Cudmore Robert, H., & Meer, J. M., Jr (2001). Map manager QTX, cross-platform software for genetic mapping. *Mammalian Genome*, *12*, 930–932.
- Martin, G. B., Williams, J. G. K., & Tanksley, S. D. (1991). Rapid identification of markers linked to a *Pseudomonas* resistance gene in tomato by using random primers and near-isogenic lines. *Proceedings of the National Academy of Sciences of the United States of America*, *88*, 2336–2340.
- Mateu-Andres, I., & De Paco, L. (2005). Allozymic differentiation of the *Antirrhinum majus* and *A. siculum* species groups. *Annals of Botany*, *95*(3), 465–473.
- McCouch, S. R., & Doerge, R. W. (1995). QTL mapping in rice. *Trends in Genetics*, *11*, 482–487.
- Meudt, H. M., & Clarke, A. C. (2007). Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends in Plant Science*, *3*, 1360–1385.
- Milbourne, D., Meyer, R. C., Collins, A. J., Ramsay, L. D., Gebhardt, D., et al. (1998). Isolation, characterisation and mapping of simple sequence repeat loci in potato. *Molecular and General Genetics*, *259*, 233–245.
- Mohan, M., Nair, S., Bhagwat, A., et al. (1997). Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding*, *3*(2), 87–103.
- Morgante, M., & Olivieri, A. M. (1993). PCR-amplified microsatellites as markers in plant genetics. *Plant Journal*, *3*, 175–182.
- Moreau, L., Charcosset, A., Hospital, F., & Gallais, A. (1998). Marker-assisted selection efficiency in populations of finite size. *Genetics*, *148*, 1353–1365.
- Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., et al. (2018). DNA molecular markers in plant breeding: Current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment*, *32*, 261–285.
- Nybom, H., Rogstad, S. H., & Schaal, B. A. (1990). Genetic variation detected by use of the M13 DNA fingerprint probe in *Malus*, *Prunus*, and *Rubus* (Rosaceae). *Theoretical and Applied Genetics*, *79*, 153–156.
- Orr, H. T., & Zoghbi, H. Y. (2007). Trinucleotide repeat disorders. *Annual Review of Neuroscience*, *30*, 575–621.
- Pammi, S., Schertz, K., Xu, G., Hart, G., & Mullet, J. E. (1994). Random amplified polymorphic DNA markers in sorghum. *Theoretical and Applied Genetics*, *89*, 80–88.
- Paran, I., & Michelmore, R. W. (1993). Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. *Theoretical and Applied Genetics*, *85*, 985–993.
- Paterson, A. H. (1996). Making genetic maps, Academic Press, Austin, Texas. In A. H. Paterson (Ed.), *Genome Mapping in Plants* (pp. 23–39). San Diego, California: R.G. Landes Company.
- Patishtan, J., Hartley, T. N., Fonseca de Carvalho, R., & Maathuis, F. J. (2018). Genome-wide association studies to identify rice salt-tolerance markers. *Plant, Cell & Environment*, *41*, 970–982.
- Pearson, C. E., Edamura, N. K., & Cleary, J. D. (2005). Repeat instability: Mechanisms of dynamic mutations. *Nature Reviews Genetics*, *6*, 729–742.
- Powell, W., et al. (1996). Polymorphism revealed by simple sequence repeats. *Trends in Plant Science*, *1*, 215–222.
- Price, A. H. (2006). Believe it or not, QTLs are accurate!. *Trends in Plant Science*, *11*, 213–216.
- Reddy, M.P., Sarla, N., Neeraja, C.N., & Siddiq, E.A. (2000) Assessing genetic variation among Asian A-genome *Oryza* species using inter simple sequence repeat (ISSR) polymorphism. Fourth International Rice Genetics Symposium, IRRI, Philippines. Abstracts. p. 212.
- Resende, R. T., et al. (2018). Genome-wide association and regional heritability mapping of plant architecture, lodging and productivity in *Phaseolus vulgaris*. *G3: Genes, Genomes, Genetics*, *8*, 2841–2854.
- Risch, N. (1992). Genetic linkage: interpreting lod scores. *Science*, *803*–804.
- Robinson J.P., & Harris S.A. (1999). Amplified fragment length polymorphisms and microsatellites: A phylogenetic perspective. In “Which DNA Marker for Which Purpose? Final Compendium of the Research Project Development, Optimisation and Validation of Molecular Tools for

- Assessment of Biodiversity in Forest Trees in the European Union DGXII Biotechnology FW IV Research Programme Molecular Tools for Biodiversity” (E. M. Gillet, ed.).
- Ruperao P., & Edwards D. (2014). Bioinformatics: Identification of Markers from Next- Generation Sequence Data Plant Genotyping; pp 29–47.
- Russell, J. R., Fuller, J. D., Macaulay, M., Hatz, B. G., Jahoor, A., et al. (1997). Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *Theoretical and Applied Genetics*, *95*, 714–722.
- Sarin, B., Mohanty, A., & demente, J. P. M. (2013). PCR-RFLP to distinguish three *Phyllanthus* sp, commonly used in herbal medicines. *South African Journal of Botany*, *88*, 455–458.
- Sehgal, D., Singh, R., & Rajpal, V. R. (2016). Quantitative trait loci mapping in plants: concepts and approaches. In V. Rajpal, S. Rao, & S. Raina (Eds.), *Molecular breeding for sustainable crop improvement* (pp. 31–59). Cham: Springer International.
- Semagn, K., Bjornstad, A., & Ndjiondjop, M. N. (2006). An overview of molecular marker methods for plants. *African Journal of Biotechnology*, *5*, 2540–2568.
- Sharma, P. N., Díaz, L. M., & Blair, M. W. (2013). Genetic diversity of two Indian common bean germplasm collections based on morphological and microsatellite markers. *Plant Genetic Resources*, *11*(2), 121–130.
- Smith, J. S. C., & Williams, J. G. K. (1994). Arbitrary primer mediated fingerprinting in plants: case studies in plant breeding, taxonomy and phylogeny. In B. Schierwater, B. Streit, G. P. Wagner, & R. DeSalle (Eds.), *Molecular ecology and evolution: Approaches and applications* (pp. 5–15). Switzerland: Birkhauser Verlag Basel.
- Soller, M., & Beckmann, J. S. (1983). Genetic polymorphism in varietal identification and genetic improvement. *Theoretical and Applied Genetics*, *67*, 25–33.
- Soranzo, N., Provan, J., & Powell, W. (1999). An example of microsatellite length variation in the mitochondrial genome of conifers. *Genome*, *42*, 158–161.
- Staub, J. C., Serquen, F. C., & McCreight, J. A. (1997). Genetic diversity in cucumber (*Cucumis sativus* L.). An evaluation of Indian germplasm. *Genetic Resources and Crop Evolution*, *44*(4), 315–326.
- Tang, J., Vosman, B., Voorrips, R. E., van der Linden, C. G., & Leunissen, J. A. (2006). QualitySNP: A pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics*, *7*, 438.
- Tanksley, S. D. (1993). Mapping polygenes. *Annual Review of Genetics*, *27*, 205–233.
- Tanksley, S. D., Bernatzky, R., Lapitan, N., & Prince, J. P. (1988). Conservation of gene repertoire but not gene order in pepper and tomato. *Proceedings of the National Academy of Sciences of the United States of America*, *85*, 6419–6423.
- Tautz, D., & Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research*, *12*, 4127–4138.
- Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development of cDNA derived microsatellite markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*, *106*, 411–422.
- Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding and Biotechnology*, *2*, 195–212.
- Torress, A. M., Weeden, N. F., & Martin, A. (1993). Linkage among isozyme, RFLP, and RAPD markers. *Plant Physiology*, *101*, 394–352.
- Tsumura, Y., Ohba, K., & Strauss, S. H. (1996). Diversity and inheritance of inter-simple sequence repeat polymorphisms in Douglas-fir (*Pseudotsuga menziesii*) and sugi (*Cryptomeria japonica*). *Theoretical and Applied Genetics*, *92*(1), 40–45.
- Utz, H., & Melchinger, A. (1996). PLABQTL: A program for composite interval mapping of QTL. *Journal of Quantitative Trait Loci*, *2*(1).
- Vejl, P. (1997). Identification of genotypes in hop (*Humulus lupulus* L.) by RAPD analysis using program Gel Manager for Windows. *Rostlinna Vyroba*, *43*, 325–331.
- Voet, D., & Voet, J. G. (2004). *Biochemistry* (3rd ed.). John Wiley and Sons Inc.
- Wang, M. L., Barkley, N. A., & Jenkins, T. M. (2009). Microsatellite markers in plants and insects. Part I. Applications of biotechnology. *Genes Genomes Genomics*, *3*, 54–67.
- Wang, Z. Y., & Tanksley, S. D. (1989). Restriction fragment length polymorphism in *Oryza sativa* L. *Genome*, *32*, 1113–1118.
- Weising, K., & Gardner, R. C. (1999). A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. *Genome*, *42*, 9–19.
- Weller, J. W., & Reddy, A. (1997). Fluorescent detection and analysis of RAPD amplicons using the ABI PRISM DNA sequencers. In M. R. Micheli, & R. Bova (Eds.), *Fingerprinting methods based on arbitrarily primed PCR* (pp. 81–92). Springer Lab Manual.
- Wenzl, P., Carling, J., Kudrna, D., et al. (2004). Diversity arrays technology (DArT) for whole-genome profiling of barley. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9915–9920.
- Xu, Y. (2010). *Molecular plant breeding*. Wallingford: CABI.
- Young, N. D. (1996). QTL mapping and quantitative disease resistance in plants. *Annual Review of Phytopathology*, *34*, 479–501.
- Yu, J. K., Dake, T. M., Singh, S., Bensch, D., Li, W., et al. (2004). Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. *Genome*, *47*, 805–818.
- Zane, L., Bargelloni, L., & Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*, *2002*(11), 1–16.
- Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, *136*, 1457–1468.
- Zhang, J., Wheeler, D. A., Yakub, I., Wei, S., Sood, R., Rowe, W., Liu, P. P., et al. (2005). SNPdetector: A software tool for sensitive and accurate SNP detection. *PLOS Computational Biology*, *1*, e53.
- Zhang, P., Zhong, K., Shahid, M. Q., et al. (2016). Association analysis in rice: From application to utilization. *Frontiers in Plant Science*, *7*, 1202.

- Zhang, X., & Yuan, Y. (2019). Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize. *Frontiers in Plant Science*, *9*, 1919.
- Zhang, Y., Wang, L., Xin, H., Li, D., Ma, C., Ding, X., et al. (2013). Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. *BMC Plant Biology*, *13*, 141.
- Zietkiewicz, E., Rafalski, A., & Labuda, D. (1994). Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics*, *20*(2), 176–183.

This page intentionally left blank

Deciphering comparative and structural variation that regulates abiotic stress response

Zeba Seraj¹, Sabrina Elias², Saima Shahid³, Taslima Haque⁴, Richard Malo⁵ and Mohammad Umer Sharif Shohan¹

¹Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka, Bangladesh, ²Department of Life Sciences, Independent University Bangladesh, Dhaka, Bangladesh, ³Donald Danforth Plant Science Center, St. Louis, MO, United States, ⁴Department of Integrative Biology, University of Texas at Austin, Austin, TX, United States, ⁵Life Science Division, Overseas Marketing Corporation Pvt. Ltd., Dhaka, Bangladesh

33.1 Introduction

Abiotic stresses on plants are estimated to currently cause a loss of 50% in crop productivity (Pandey, Irulappan, Bagavathiannan, & Senthil-Kumar, 2017). Further increase in temperatures due to climate change is expected to exacerbate heat and drought stress as well as enhance salinity in soils due to evaporation. Salinity is already a major problem in many coastal areas due to sea level rise (Gopalakrishnan, Hasan, Haque, Jayasinghe, & Kumar, 2019). To cope with these stresses, plants switch from their regular developmental program to an altered metabolism at the expense of their reproductive potential (Annacondia, Magerøy, & Martinez, 2018). Response to abiotic stresses are not only organ or tissue-specific, these depend on the developmental stage as well, with panicle-bearing being more sensitive than seedling growth, particularly in cereals like rice (Gray & Brady, 2016; Razzaque et al., 2017; Razzaque et al., 2019). In the struggle for survival, some plants have evolved to tolerate stress (Mickelbart, Hasegawa, & Bailey-Serres, 2015). The evolved genotypes within the same species are referred to as landraces and can be used as donors for tolerance traits, provided that their mechanism for survival is understood in finer detail. Often a multitude of coordinated activities can be responsible for tolerance. For instance in case of salinity tolerance in rice, some of the major mechanisms are cell expansion (Jadamba, Kang, Paek, Lee, & Yoo, 2020), retrograde signaling (Huang et al., 2020), Na⁺ transport and extrusion (Srikantharajah et al., 2020), G-protein signaling as well as change in membrane potential (Razzaque et al., 2017), shape alteration in mesophyll chloroplast to allow greater dissipation of light energy (Oi et al., 2019), maintenance of low reactive oxygen species (Bhattacharjee, 2012) and induction of RNA chaperones (Ganie, 2020). Not all of these mechanisms is present in any particular landrace and expression quantitative trait loci (eQTL) are a good method of determining the relative importance of a specific mechanism and the genes regulating it. eQTLs combine mapped genetic loci with RNA seq gene expression studies (Guo et al., 2019).

Each of the mechanisms for fighting stress may be regulated and fine-tuned by small RNA-mediated regulation of transcripts (Goswami, Mittal, Gautam, Sopory, & Sanan-Mishra, 2020), alternative splicing, (Fu, Shen, Kuang, Wu, & Zhang, 2019) or reprogramming at a deeper level in the genome such as methylation of chromatin histones or DNA. Methylation has been reported for defense against salinity stress (Chen, Luo, Wang, & Wu, 2010) and against drought stress (van Dijk et al., 2010). Even the number of stomata can be altered when cells sense reduction in relative humidity and this is controlled by the RNA-directed DNA methylation (RdDM) pathway (Tricker, Gibbings, Rodríguez López, Hadley, & Wilkinson, 2012). Drought causes genome-wide changes at the cytosine methylation level (Colaneri & Jones, 2013). Heat stress has been reported to induce the de-condensation of rDNA loci in rice and Arabidopsis (Pecinka et al., 2010; Tomás, Brazao, Viegas, & Silva, 2013). Epigenetic control of transposable element expression and transposition may also result in additional control (Lisch, 2013; Wang, Weigel, & Smith, 2013). In recent years, there has been a technological watershed not only for developing efficient methods to dissect these stress regulatory

pathways, but also in the area of computational biology for systems-level analysis of the interconnected datasets. The former encompasses high throughput methods to generate multiomics data such as total RNA, direct sequencing of native RNAs, small RNA, RNA degradome and Methyl-C sequencing and the latter covers efficient ways to process and robust statistical methods to analyze these Big data (Sedlazeck, Lee, Darby, & Schatz, 2018; Simon et al., 2009; Yang et al., 2020).

In order to avail all the above information into use for crop improvement programs, it is essential that there is targeted characterization of germplasm in seedbanks. Genome Wide Association Study (GWAS) studies in cereals like rice and other cereals has already been reported for plant architectural and grain quality traits (Dwivedi, Scheben, Edwards, Spillane, & Ortiz, 2017; Huang et al., 2010). Some of these have been colocalized with known quantitative trait loci (QTLs) for desired phenotypes (Biscarini et al., 2016). SNPs and germplasm-based GWAS for phenology and yield in legumes has also been reported (Dwivedi et al., 2017). It is also essential that the breeder, the physiologist, the molecular biologist and the bioinformatician work in a coordinated manner in order to produce crops to feed our future generations. It will however help the biological scientists to acquire some computational skills for data-intensive aspects of enabling plants to survive and thrive in our changing environments.

33.2 Expression quantitative trait loci and their functional significance

Biological information flows from DNA to RNA to protein towards visible phenotypes. Each of the steps for this, for example, transcription, translation as well as post-translational levels are regulated by specific factors. eQTL connects the sequence level polymorphism to gene expression variation. Quantitative trait loci are regions in the genome that are associated with a quantitative trait of interest, for example, plant height, yield, etc., and other traits which altogether determine the performance of a plant under particular abiotic stress/s. In eQTL analysis, the expression value of gene is also considered as a quantitative trait. Dynamism of a biological system depends on its ability to switch expression of genes in response to environmental demand. Hence differential gene expression is observed when specific genotypes encounter abiotic stress. A QTL region for a specific trait can comprise more than hundreds of genes depending on the span of the boundary markers. eQTLs can therefore not only identify genes associated with the selected molecular markers, but also pinpoint important differentially expressed genes underlying the physiological QTL (pQTL) region. Variation in temporal (time point) and spatial (tissue) gene expression as well as in multiple developmental stages under continuous abiotic stress can identify underlying mechanisms that a plant adopts to combat the encountered stress. Computational biology and statistical modeling have always been an integrated part of this comparative association study. Advancement of next generation sequencing (NGS) strategies and high throughput phenotyping facilities as well as associated large volume of biological data have made such bioinformatics an essential part of modern crop improvement schemes. In the following subsections, recent genotyping technologies for determining sequence level polymorphism have been described. High throughput techniques for identifying gene expression data and its analysis, that is commonly used in crop genomics is also explained. Expression QTL mapping and its significance are then discussed by explaining the correlation of the genotyping with the expression data.

33.2.1 Molecular marker system for genotyping

Molecular markers are like a flag in the genome that can mark variation or polymorphism in genome sequences of different cultivars within the same species. Common DNA markers are Randomly Amplified Polymorphic DNA, Restriction Fragment Length Polymorphism, Amplified Fragment Length Polymorphism, Simple Sequence Repeat (SSR), Single Nucleotide Polymorphism (SNP), Cleaved Amplified Polymorphic Sequences, etc. The simple sequence repeats (SSR) have been the marker of choice for different species for many years, where polymerase chain reaction (PCR) primers are designed to flank the repeat regions to determine the differences in repeat motifs among the genotypes being analyzed. Molecular markers in the polymorphic region between two specific cultivars can be used to track the inheritance of the marker allele in subsequent generations. Hence, in a biparental population having 2 different alleles linked to a specific trait, segregation of the markers will occur in the F₂ generation, comprising homozygous alleles of the individual parents as well as heterozygous alleles. Using single seed descent method (Pazos-Navarro, Castello, Bennett, Nichols, & Croser, 2017) advancement of generations, will result in reduction of heterozygosity if a specific allele follows the 1:2:1 ratio of the first law of Mendelian genetics. Molecular markers are inherited along with the DNA loci of the trait of interest, because they are close enough not to be segregated from each other. Linkage maps are generated from the recombination frequency of the alleles where the estimated distance between two markers is

positively correlated with their recombination frequency, that is, the higher the recombination frequency, the more distant two markers are from each other. The basic principle of mapping QTL is to track the parental allelic segregation in progenies and associate those with physiological traits using appropriate statistical methods. In order to successfully map the QTLs for a trait of interest, the two parents must show enough variation in terms of that trait and a robust marker system needs to be used. The mapping population also need to gain enough homozygosity for specific alleles (Collard, Jahufer, Brouwer, & Pang, 2005). Different mapping populations can be used for this purpose, such as, recombinant inbred lines (RILs), Near Isogenic Lines, Doubled Haploid, and F₂. F_{2:3} lines are also used, where each F₃ line derived from specific F₂s can be considered as replicates (Zhang & Xu, 2004). Thus genotyping of F₂ and Phenotyping of F₃ can also be used to map QTLs successfully without waiting for years for the generation of inbred lines (Haque et al., 2020). Phenotypic differences can be due to a few loci with great effects and are less due to large loci with smaller effects (Mäki-Tanila & Hill, 2014). Advancement in sequencing technology and completion of major cereal rice genome sequencing project accelerated the compilation of rice SSR databases (McCouch et al., 2002; Sasaki, 2005; Temnykh et al., 2000) which have helped identify many QTL regions for traits of interest including a major QTL named *Saltol* for salinity tolerance. *Saltol* was mapped on chromosome 1 using F₈ RILs of Pokkali/IR29 cross (Gregorio, Islam, Vergara, & Thirumeni, 2013) and was later found to harbor genes that can play roles in maintaining sodium to potassium ratio in shoots (Das, Nutan, Singla-Pareek, & Pareek, 2015; Ren et al., 2005).

The main disadvantage of SSR marker system is that its numbers in the genome are limited in contrast to SNP markers. Large number of polymorphic markers can help in precise tracking of alleles in subsequent generation of progenies. Moreover, advancement in NGS technologies have accelerated identification of large number of variants in the genome and the most commonly used markers are now SNPs. Polygenic traits are affected by multiple genes, but statistically significant genotype to phenotype association often helps identify specific regions that are more important for the specific trait of interest. Most widely used methods for QTL mapping are single marker analysis, simple interval mapping, composite interval mapping (CIM), etc. (Lin, Sasaki, & Yano, 1998; Liu, 2017; Tanksley, 1993). In single marker analysis, t-tests, ANOVA (Analysis Of Variance), and linear regression models are used. The phenotypic variation due to the QTL can be explained by coefficient of determination (R²) in the linear regression. Use of larger number of segregating markers covering the entire genome is essential making SNP markers ideal (Tanksley, 1993). CIM is more powerful as it combines interval mapping and linear regression. The R/qtl (Broman, Wu, Sen, & Churchill, 2003) package offers many functions like *scanone* for single QTL analysis as well as *mqmscan* for multiple QTL scanning. After QTL mapping, the plants with desired QTLs are advanced to create RILs which can be used as donor plants in Marker Assisted Back-Crossing (MABC) programs with a high-yielding variety or any variety whose genomic background is desired. Also multiple QTLs can be pyramided during MABC program by selecting the QTL marker alleles and avoiding the background marker alleles of the donor plant.

The advancement in SNP detection methods have made its use in crop genomics and improvement remarkable. Earlier, Expression Sequence Tag database or sequencing of gene amplicons using Sanger sequencing were used to identify SNPs. Resequencing of whole genomes opened a path for discovering more SNPs. For instance, Shen et al. (2004) reported identification of 1.7 million SNPs, and Feltus et al. (2004) reported 384,431 high quality SNPs by comparing the japonica rice Nipponbare and indica rice 93–11 genomes. Later, 1536 SNP array was designed from the polymorphic sites among the 5 major subpopulations of *Oryza sativa* (Zhao et al., 2010). Thomson (Thomson, 2014) reported a 384 SNP assay but with lower resolution. Most genotyping by sequencing (GBS) methods take the advantages of reducing the genome complexity using restriction enzymes coupled with the NGS technologies. Complexity Reduction of Polymorphic Sequence (CroPS) and Restriction Site Associated DNA (RAD) (Scheben, Batley, & Edwards, 2017) can filter out duplicated SNP and are computationally robust without the need for complete sequencing of all the genotypes in the population. In Double Digested RAD or ddRAD two restriction enzymes are used (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). The digestion is followed by adapter ligation, shearing, end repair, second adapter ligation, and size selection. Digested fragments are tagged and amplified using PCR and sequenced from the two generated ends. Multiple samples can be pooled and sequenced together using different barcodes. In ddRAD method one rare cutter and one frequent cutter is used while other methods can use one enzyme. Diversity Array Technology or DArT-based sequencing method uses an intelligent selection of the genome to target polymorphic regions closely associated with gene rich regions. A combination of restriction enzymes can separate low copy sequences from the repetitive fractions. Low copy sequences are most informative for marker discovery and typing. Initially DArT was used in microarray. Classic DArT markers are substituted by DArTseq (Jaccoud, Peng, Feinstein, & Kilian, 2001) markers based on GBS. For setting up DArTseq in a new organism, the first step is usually the optimization of genomic complexity reduction using restriction enzyme combinations. Sequencing can generate higher number of markers compared to the array version of DArT.

GBS has helped discover DNA markers that are polymorphic in a specific variety compared to the reference genome. Advancement in sequencing technologies has improved the SNP discovery pipeline, resulting in identification of a higher number of polymorphic markers. SNPs and indels variants can be generated by calling haplotypes in a variant call format (vcf) file. Important variant calling tools are: Genome Analysis Toolkit (GATK), mpileup function of samtools, etc. (McKenna et al., 2010; Yao et al., 2020), and can be used for this purpose. Normally a workflow includes sequence read filtering, alignment, identifying SNPs from aligned tags and scoring of all the discovered SNPs for various coverage, depth, and genotypic statistics. After identification of the SNPs from all tags, duplicates and variants are called followed by imputation if needed. Imputation refers to replacing missing data values with substituted values. Variants are filtered according to a specific allele frequency threshold. Selection of specific restriction enzymes to shear the genome to generate fragments in nonrepetitive regions and targeting low copy regions is important for nondistorted coverage of the whole genome. This also reduces the alignment problem in genetically highly diverse species (Elshire et al., 2011). QTL mapping has some limitations, for example, the segregated allelic diversity between two parents are only assayed and the recombination amount in the RIL places a limit on mapping resolution. Use of advanced intercross RIL as well as multiparent advanced generation intercross lines can however increase the allelic diversity where multiple genetically diverse accessions are intercrossed before establishing the RIL (Kover et al., 2009). GWAS overcomes the main limitation of biparental analysis as it involves diversity or accession panels with a large number of accessions (>200), landraces, or breeding materials. QTLs are identified here by using marker–trait association and the linkage disequilibrium between polymorphic markers or SNPs of the diverse set of germplasm (Zhu, Gore, Buckler, & Yu, 2008). GWAS can serve as being an informed choice of parents for QTL analysis as well as being suggestive for gene choice in functional genomics studies. Famoso et al. (2011) used both GWAS and biparental QTL mapping to elucidate genetic architecture of Aluminum tolerance in rice and by identifying several subpopulation specific QTLs they concluded that the subpopulation structure in rice has a major role in the tolerance. In an integrated analysis of GWAS, QTL mapping, and RNAseq, Guo et al. (2019) could identify 6 colocalized loci between GWAS and QTL ultimately pinpointing 44 genes responsible for seed vigor in rice. For rice, a diverse allelic resource was generated using 3,000 rice genomes by analyzing allele frequencies in rice subpopulations, thus improving the downstream analysis in the rice genome (Alexandrov et al., 2014). However, GWAS has a high false discovery rate, hence the association between markers and traits are often verified by development of parental populations and QTL mapping.

33.2.2 Transcript abundance measurement by RNA sequence

RNAseq has provided a big leap for the total transcript count and gene expression analysis in an organism (Wang, Gerstein, & Snyder, 2009) compared to the previously used microarray (Alonso-Simón et al., 2010). The microarray depends on preidentified probe sequences on a chip for hybridization. Serial analysis of gene expression was also used earlier which was a sequencing-based high throughput method for gene expression followed by massively parallel signature sequencing. Microarrays use the hybridization approach where chips containing probes for specific transcripts are attached and complementary DNA from the RNA are allowed to hybridize with the probe. The intensity of the fluorescent dye attached to the hybridized probe can indicate the expression intensity of the gene. In RNAseq studies, the number of reads mapped to the total number of genes, that is, the transcript counts in an organism are considered as indicator for the level of expression. Up or down regulation of the same gene under nonstress and stress condition in tolerant or sensitive cultivars or in a population can indicate potential candidates involved in conferring salt tolerance. A common pipeline is to clean the transcript count raw data with Trimmomatic (Bolger, Lohse, & Usadel, 2014), followed by mapping with a reference genome/transcriptome. The mapping step needs high computational power. Reads which span exon junction can map to multiple location making the mapping step complicated. For RNAseq, illumina's HiSeq system with 150 bp paired end reads with sufficient coverage is the most popular method. An alignment tool called STAR (Dobin et al., 2013) is commonly used for ultrafast mapping of the transcripts to the reference genome. The read counts then need to be measured, for example, using featureCounts (Liao, Smyth, & Shi, 2014). The principle that is followed is: the more the mapped read counts for a specific gene, the more the transcript number and expression level.

Reduction in cost for multiple sample sequencing can be achieved by adopting different modified RNA sequencing approaches, for example, a 3' tag based RNAseq method by Meyer, Aglyamova, and Matz (2011) along with multiplexing samples using barcodes. Rarefaction or saturation curves can be plotted to evaluate whether the sequencing depth is enough to represent all transcripts, which can discern whether collecting more data will actually be able to detect more transcripts. Differential expression analysis can be performed using R packages like DESeq2 (Love, Huber, & Anders, 2014), edgeR (Robinson, McCarthy, & Smyth, 2010), etc., via Bioconductor. This helps elucidate whether there is a significant difference in the mean expression levels of different sample groups (e.g., stress vs nonstress). In general,

DESeq2 normalizes the raw counts by removing the library depth bias followed by estimation of gene-wise dispersion. Negative binomial models are then fitted followed by hypothesis testing using Wald test or Likelihood ratio test (Love et al., 2014). The edgeR package uses Generalized Linear models and relates the linear regression to the response variable (Robinson et al., 2010). Other tools like JMP genomics (SAS Institute Inc, Cary, NC) use different models for normalization (such as Kernel Density Mean of M component). Differential expression analysis has already shed light on groups of genes that are upregulated or downregulated under salinity stress in different time points and different tissues (Razzaque et al., 2017; Razzaque et al., 2019; Wang et al., 2018; Yeo, Bhavé, & San Hwang, 2018).

33.2.3 Connecting genomic variation to expression variation

A merger of genomics and genetics was proposed by Jansen and Nap (Jansen & Nap, 2001) also known as eQTL. As mentioned earlier, in eQTL mapping the transcript abundance of each gene is considered as a quantitative trait. Combining the transcript count information with the genotyping information or linkage map generated from the molecular markers can quantify and perform a multifactorial dissection of the RNA into its underlying genetic components or mapping positions. Cis-eQTLs indicate which variation in the gene expression map to the SNPs of the genes themselves and trans-eQTLs indicate which variation maps to distant genome locations.

RNAseq reads can also be directly aligned to the respective references to call the variants followed by haplotype imputation. Recently Galpaz et al. (2018) have performed SNP calling by analyzing the RNAseq analysis in Melon for fruit quality-related traits. A melon consensus linkage map merging data from eight populations including 414 × Dul was constructed with 1,592 markers. Several causative genes were mapped to single gene resolution and unknown genes that affect fruit aroma and flesh color were identified and functionally or genetically validated. In another study by Li et al. (2018), fatty acid composition, flowering time and growth trait-related QTLs and eQTLs were identified in the allopolyploid *Brassica napus*. This study used BradSeq library construction method where 3' digital gene expression libraries were extended to full transcript coverage in a shotgun type strand-specific approach. The authors used the RNAseq data for genotyping using the R package onemap (Margarido, Souza, & Garcia, 2007) with the default value of LOD score 3 and maximum recombination fraction 0.5 (Li et al., 2018). For QTL mapping they used both *scanone* and CIM. For eQTL analysis they used the multiple algorithm of interval mapping method in R/qtl package (Broman et al., 2003). It is important to note that proper statistical models are needed to retrieve the information from RNAseq or microarray data and associating these with the linkage map. Tools like R/eQTL R/qTL, eMAP, Merlin, FastMap, Matrix eQTL are among those being used, where the last one shows the fastest performance (Shabalina, 2012).

An eQTL study on rice shoots at 72 h after germination from 110 RILs of Zhenshan97 and Minghui 63 cross could identify 26,051 eQTLs and 171 eQTL hotspots under nonstress condition (Wang et al., 2010). Specifically, eQTLs for e-traits (expression as a trait) that were involved in DNA metabolic process were significantly enriched in the eQTL hotspots on chromosome 3, 5, and 10. This study also found correlation between shoot dry weight QTLs and eQTLs revealing potential candidate genes for the phenotypic trait (Wang et al., 2010). Comparative transcriptomics of salt-tolerant and sensitive rice genotypes in a different study showed that under salt stress more genes were downregulated at 48 h in both genotypes, but at 72 h the numbers of upregulated and downregulated genes were almost equal (Wang et al., 2018). Differential gene expression of rice breeding lines has also been investigated under salt stress in both seedling (Razzaque et al., 2019) and reproductive stages (Razzaque et al., 2017) in both shoot and root tissue under two time points. These types of multifactorial models are beneficial in elucidation of tolerance mechanism and can identify developmental stage-specific temporal and spatial eQTLs from the same linkage map compared to their nonstress expression profile.

In case of eQTL analysis, the expression values are considered as quantitative traits and this is a useful strategy for exploring the regulatory relationship between genes (Kliebenstein, 2009). Certain locations in the genome can act as hotspots (above a specific threshold value) by regulating multiple gene expression. Mostly these appear as trans-eQTL and they can work together in specific biological pathways to regulate certain phenotypic traits and can form functional networks of correlated genes. Different coexpression network construction method like Mutual ranking, Weighed Gene Coexpression Network Analysis can be performed on these genes to elucidate their relevance to specific biological pathways. Hence combining the eQTL information with coexpression studies can identify regulatory candidates and increase resolution of the analysis. Baker et al. (2019) characterized the mechanistic connections between genomic architecture, gene expression networks and phenotypic variation throughout plant development using targeted eQTL analysis.

eQTL studies can bypass the positional cloning process (Hansen, Halkier, & Kliebenstein, 2008) by narrowing down specific genes underlying a phenotypic QTL region. They can be considered as a sequence-based genetic framework

map identifying genes associated with the stress being studied. The major challenge in analyzing eQTL is the complexity of genome-wide gene expression data. There are thousands of genes in a species and when all of them are considered as individual traits, it becomes computationally intensive to calculate the association. RNAseq and eQTL based approaches are also challenging when used with crop genomes with more ploidy. To address such challenges, several guidelines for eQTL mapping in allopolyploid organisms have been proposed recently (Fan, Devos, & Schliekelman, 2020). Additionally, improved computational and statistical power can help in more precise mapping of eQTLs.

Besides the physiological QTLs, merging GWAS and eQTL data is also of much benefit. This was demonstrated by Li et al. (2020), who pinpointed a specific gene encoding KIP-related protein to be a master regulator of the genes responsible for cell wall synthesis contributing to fiber length in cotton. eQTLs can map the changes in the expression level of a gene to its structural variations even under stress conditions. For instance, eQTL mapping identified that *trans-regulatory elements (TRES)* and transcription factor binding site evolution are the key players in drought response of C4 perennial grass *Panicum hallii* (Lovell et al., 2018). Thus the genome-wide allelic expression differences under both nonstress and stress condition can shed light on environmental perturbations and how variation in regulatory elements shapes phenotypic diversity.

33.3 Regulatory small RNAs

Plants express a diverse range of noncoding small RNAs (20–24 nt) that mediate posttranscriptional or transcriptional regulation of gene expression (for review, see Axtell, 2013). Regulatory plant small RNAs can be broadly divided into two categories based on their biogenesis patterns—microRNAs (miRNAs) and small interfering RNAs (siRNAs). miRNAs are derived from single-stranded hairpin precursors and generally function in posttranscriptional regulation of messenger RNAs through target cleavage or translational inhibition. miRNAs regulate a wide range of biological processes in plants including development, growth, biotic, and abiotic stress response (reviewed in Jones-Rhoades, Bartel, & Bartel, 2006; Martin, Liu, Goloviznina, & Nonogaki, 2010; Song, Li, Cao, & Qi, 2019).

In contrast to miRNAs, siRNAs are derived from long double-stranded RNA precursors. Plant siRNAs can be further divided into several subcategories such as heterochromatic siRNAs (hc-siRNAs), phased siRNAs (phasiRNAs), and *trans*-acting siRNAs (tasiRNAs). Hc-siRNAs (usually 24 nucleotides long) account for the majority of the expressed plant small RNAs. Hc-siRNAs function in transcriptional silencing of transposons and repetitive elements in the genome through the RdDM pathway (Matzke, Kanno, & Matzke, 2015; Zhang, Lang, & Zhu, 2018). PhasiRNAs and tasiRNAs are both secondary siRNAs whose biogenesis is triggered by an initial miRNA or siRNA-directed cleavage of target RNA (Axtell, 2013; Liu, Teng, Xia, & Meyers, 2020).

33.3.1 Discovery and annotation of small RNAs based on deep sequencing

Early discoveries of plant small RNAs involved Sanger sequencing of cloned products (Llave, Kasschau, Rector, & Carrington, 2002; Reinhart, Weinstein, Rhoades, Bartel, & Bartel, 2002) and computational prediction of conserved miRNAs across different species (Zhang, Pan, Wang, George, & Anderson, 2005). However, such approaches limited the discovery of species- and lineage-specific small RNAs. The development of high-throughput sequencing technologies (Fahlgren et al., 2007; Fahlgren et al., 2009) and improved genome assemblies greatly accelerated large-scale systematic analysis of small RNAs in diverse plant species (Montes et al., 2014; You et al., 2017). Concomitantly, genetic studies characterizing structural determinants for miRNA and siRNA biogenesis defined community standards for small RNA loci annotation, especially for miRNAs (Meyers et al., 2008). As a result, a large number of web-based and stand-alone tools dedicated to miRNA discovery have been developed over the years (Kuang, Wang, Li, & Yang, 2019; Mohorianu, Stocks, Applegate, Folkes, & Moulton, 2017; Morgado & Johannes, 2019; Tseng et al., 2018). However, variations in the quality and stringency of annotation practices have also led to many spurious annotations (Coruh, Shahid, & Axtell, 2014; Kozomara & Griffiths-Jones, 2014; Taylor, Tarver, Foroozani, & Donoghue, 2017). To address these discrepancies, a revised set of guidelines have been recently proposed for improved annotation of plant small RNAs (Axtell & Meyers, 2018; Kozomara & Griffiths-Jones, 2014). Despite these efforts, relatively fewer tools are available for annotating siRNAs and other classes of small RNAs (Axtell, 2013; Mohorianu et al., 2017; Shahid & Axtell, 2013). Furthermore, as the majority of the siRNAs are derived from transposable elements (TE) and repeats, these often map to multiple regions of the genome. Thus the placement of multimapping reads can greatly affect the discovery of siRNA loci (Johnson, Yeoh, Coruh, & Axtell, 2016). Accurate assembly and annotation of repetitive, transposable element-containing regions of the genome are also necessary for reducing false-positives in siRNA loci discovery.

Once small RNA loci have been annotated, these can be utilized for multiple downstream analyses for further characterizing their function (Fig. 33.1). Deep sequencing allows capturing genome-wide changes in small RNA abundance in different tissue samples or conditions. By using tools such as DESeq2 (Love et al., 2014) or edgeR (Robinson et al., 2010), differentially expressed small RNA loci can be easily identified from such abundance datasets. Differential expression analysis of small RNA loci has been successfully utilized by many studies to identify not only stress-responsive small RNAs but also signaling pathways important in different stress conditions. For instance, miRNA expression profiling of *Arabidopsis* seedlings unveiled cross-talk of stress signaling pathways involved in both drought and salt stress response (Barciszewska-Pacak et al., 2015).

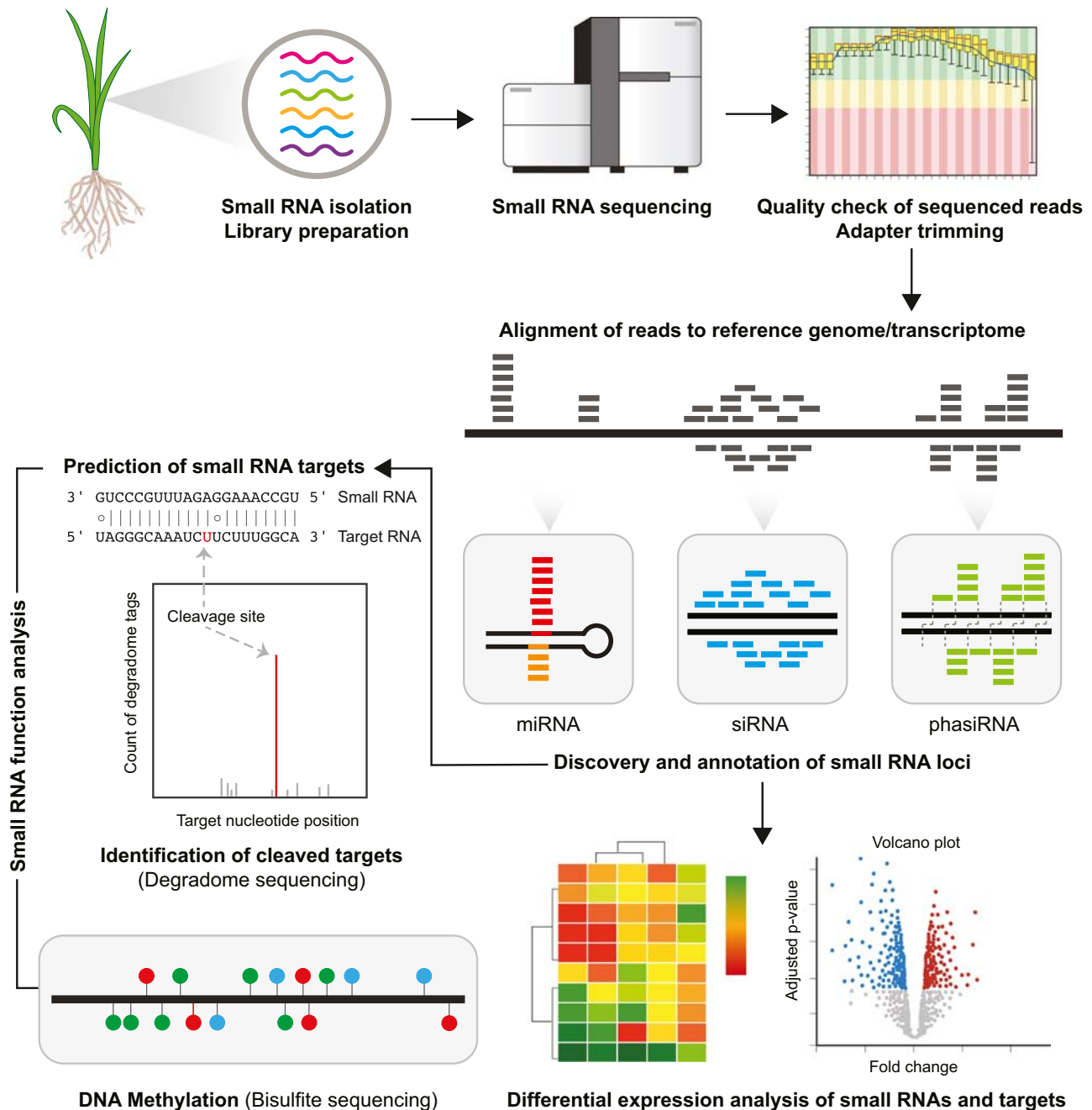


FIGURE 33.1 Bioinformatic approaches for discovery and analysis of regulatory small RNAs and their targets. (Figure created using BioRender).

33.3.2 Detection of small RNA targets

Targets of small RNAs can be predicted based on sequence complementarity. Unlike animals, plant small RNAs usually require near perfect matches with their targeted transcripts. Various tools have been developed for prediction of miRNA targets in the transcriptome (Addo-Quaye, Miller, & Axtell, 2009; Dai, Zhuang, & Zhao, 2018; Fahlgren & Carrington, 2010). Most of these tools utilize a scoring system based on the number of matches, mismatches, and G:U wobbles between miRNA and the aligned mRNA for predicting targets. Since plant miRNAs are known to cleave targets in a precise manner, cleavage of predicted targets can be validated using techniques such as degradome and PARE sequencing (Simon et al., 2009) (Fig. 33.1). Software such as CleaveLand (Addo-Quaye et al., 2009), sPARTA (Kakrana, Hammond, Patel, Nakano, & Meyers, 2014), and PAREsnip2 (Thody et al., 2018) are equipped to predict both small RNA targets and detect cleaved targets based on degradome and PARE sequencing datasets. Besides target cleavage, miRNAs are also capable of translational repression of targeted mRNAs. Integrative analysis of small RNA and their target expression profiles along with degradome sequencing allow high resolution analysis of small RNA-target interactions (Fig. 33.1). This type of integrative analysis has been successfully applied for identifying functional miRNA-target pairs in diverse processes for both model plants (Thatcher, Burd, Wright, Lers, & Green, 2015) and non-model plants (Cheng et al., 2020).

Heterochromatic siRNAs that trigger DNA methylation via RdDM pathway recognize noncoding scaffold RNAs transcribed by the plant-specific RNA Polymerase V (Pol V) (Wendte & Pikaard, 2017). In this case target recognition leads to recruitment of additional factors which mediate methylation of the associated DNA, leading to transcriptional silencing. Techniques for detecting DNA methylation include bisulfite sequencing (Plongthongkum, Diep, & Zhang, 2014), which has been discussed in detail in Section 33.4.1. Additionally, DNA methylation can be detected from direct sequencing of genomic DNA using long read technologies such as Oxford nanopore and PacBio (Flusberg et al., 2010; Laszlo et al., 2013). Long read-based techniques do not require bisulfite-converted DNA for detecting methylation and provide better coverage of repetitive and transposon-rich regions compared to short-read sequencing. Several studies have utilized such technologies to identify genome-wide DNA methylation patterns (Ni et al., 2021; Simpson et al., 2017; Tse et al., 2021). However, further development of tools and algorithms for accurate detection of modified bases from long read sequencing datasets is necessary.

33.3.3 Natural variation in small RNAs and their targets

In general, plant miRNAs and their target sites are considered to be under strong purifying selection (Ehrenreich & Purugganan, 2008; Wang et al., 2010). However, natural variations in miRNAs, miRNA precursors, and target sites have been reported in several species (de Meaux, Hu, Tartler, & Goebel, 2008; Liu, Wang, Zhu, Hu, & Sun, 2013; Liu et al., 2016; Wang et al., 2010). In some cases, such variation has been linked to phenotypic diversity. For instance, sequence polymorphism in the Arabidopsis ath-miR164 precursor affects miRNA expression level and contributes to variation in leaf shape and short architecture (Tedesco et al., 2012). In rice, a GG/AA polymorphism in the precursor of osa-miR2923a has been linked to grain length (Wang et al., 2013), while variation in the polyadenylation tail of osa-miR156h precursor affected grain yield (Zhao et al., 2015). Sequence polymorphism in upstream regulatory regions of the miRNA locus can also affect its expression. Such polymorphisms have been speculated to be the cause of variable miR397 expression linked to domestication-related phenotypes in *indica* rice (Swetha et al., 2018). Finally, polymorphism in miRNA target sites has been linked to phenotypic diversity (Duan et al., 2015; Jiao et al., 2010; Miura et al., 2010; Nair et al., 2010). These studies highlight the importance of exploring natural variations in small RNAs and their targets and provide directions for manipulating such variations for crop improvement.

33.3.4 Integrating small RNA sequencing with quantitative trait loci mapping

Although several studies have reported genome-wide expression variation in miRNA and siRNA loci across closely related species (Ma, Coruh, & Axtell, 2010; Wen et al., 2016), the underlying molecular mechanism and consequences of such variation have not been extensively examined. Deep sequencing of miRNAs in six rice accessions revealed significant miRNA expression variation between rice grains and seedling (Wen et al., 2016). This is consistent with the regulatory roles of miRNAs in development. Interestingly, several rice QTLs that affect yield-related traits such as grain number and grain weight are associated with miRNAs (reviewed in Peng, Teotia, Tang, & Zhao, 2019). In maize, mapping miRNAs to QTLs led to the discovery of several miRNA genes that colocalized with QTLs linked to waterlogging tolerance (Osman et al., 2013). Therefore combining small RNA sequencing with QTL mapping can greatly enhance

the discovery of traits affected by small RNA expression variation. This approach has been recently applied for identifying miRNA expression-related QTLs (miR-eQTLs) in several crops (Chen et al., 2020; Liu et al., 2017). For instance, Chen et al. combined miRNA expression variation in 200 maize lines along with the maize HapMap to identify four miR-eQTLs (Chen et al., 2020). As many miRNAs and siRNAs are differentially expressed in response to abiotic stress (Borsani, Zhu, Verslues, Sunkar, & Zhu, 2005; Furini, Koncz, Salamini, & Bartels, 1997; Zhang, 2015), small RNA-eQTL analysis can be also useful for determining traits linked to stress tolerance. Indeed, miRNA expression profiling in the salt-sensitive rice Pusa Basmati and salt-tolerant rice Pokkali identified several potential miR-eQTLs associated with salt tolerance (Goswami et al., 2020). Compared to miRNAs, very few studies have explored expression variation in siRNAs. However, as majority of the expressed plant small RNAs are actually siRNAs, eQTL analyses for these could broaden our understanding on their regulatory potential. A relevant example is the role of siRNA expression variation in regulating rice hybrid vigor (Zhang et al., 2014). Further development of siRNA annotation methodologies will be crucial for expanding small RNA-eQTL analysis beyond miRNAs.

33.4 Epigenomic regulation of gene expression in plant

The elements that regulate transcription of genes may reside in coding or noncoding parts of the genome and are broadly classified into two categories: cis-regulatory elements (CREs) and TREs. CREs reside in noncoding regions of the genome whereas TREs reside in the distal coding regions and can code for transcription factors (TFs), noncoding RNAs or signaling molecules. CREs are noncoding conserved DNA sequences of ~5–20 bp which often contain binding sites for different TFs (Rombauts et al., 2003). These elements may reside at the core promoter or UTR region of the target gene and are often called proximal CREs. On the other hand, distal CREs reside far away from the target gene and can come close during the process of transcription due the conformational change of the chromosome. Unfortunately, we still have very limited knowledge of plant CREs and their regulation partially due to the fact that the identification of genome-wide CREs is difficult. Since these elements have less universal sequence conservation and shorter length, their identification is analogous to finding a needle in a haystack. Active CREs can initiate, facilitate, enhance or repress transcription of the associated target genes and therefore play a chief role in the Gene Regulatory Network (GRN) of an organism. The dynamic nature of transcriptional regulation through CREs provides high flexibility for rewiring GRNs in a spatial and temporal manner. This rewiring of GRN offers a great deal of plasticity for a multicellular organism to respond differently to the changing environment. What are the ways to regulate CREs which modulates the transcription of the associated target genes? DNA Methylation, histone modification or ATP-dependent chromatin remodeling are the major regulatory mechanisms that have been studied for CRE activation or repression. These regulations are often called epigenomic regulations. The term epigenome is formed from the Greek word “epi” which means “above.” Since these epigenomic marks modify the genome and its expression without altering the genetic code hence these are called epigenomic regulations.

33.4.1 DNA methylation and its role in transcriptional regulation

DNA methylation is a conserved mechanism for eukaryotes that plays an important role not only for gene regulation but also for TE silencing, imprinting and whole chromosome inactivation. Active TEs can insert themselves into the regulatory or coding regions of the genome and can cause functional changes of regulatory elements and genes. This may cause genome instability and sometimes genetic disorders such as Hemophilia A in humans (Kazazian et al., 1988). Therefore transcriptional silencing of TEs is a key epigenomic regulation that is required to suppress the transcription of TEs. Gene imprinting and whole chromosome inactivation are the processes of epigenomic modification mostly carried out by DNA and histone methylation that controls the expression of one single gene referred to as “imprinting” or whole chromosome for the latter process. DNA methylation in eukaryotes usually occurs at the fifth position of the cytosine bases but the context varies between mammalian and plant systems. For plants, methylation can occur at CG, CHG, CHH (H = A, C, or T), whereas mammalian DNA methylation is restricted to only to the CG context (Henderson & Jacobsen, 2007). In plants several different methyl transferases have been classified: Methyltransferase 1 (MET1), Domain Rearranged Methyltransferase 2 (DMR2), Chromomethylase 2 (CMT2) and Chromomethylase 3 (CMT3) (Zhong et al., 2021). MET1 is responsible for maintenance of CG methylation during replication whereas CMT3 is the plant specific DNA methyltransferase to maintain CHG methylation (Law & Jacobsen, 2010). CHH methylation is maintained by DRM2 through the RdDM pathway and CMT2. *De novo* methylation is also mediated by the RdDM pathway which relies on DNA-dependent RNA polymerases, Pol IV, and Pol V (Gallego-Bartolomé et al., 2019).

Molecular functions of DNA methylation for the different structural features of the genome are complex and still not well-understood. DNA methylation in all contexts in the heterochromatin region, such as for TEs, is involved in silencing these elements and works as a defense system (Wang & Baulcombe, 2020) of the genome. Protein-coding genes located in euchromatin region can also be subjected to DNA methylation although the patterns of methylation can alter the fate of the target gene expression. Methylation on the coding region of a gene is often referred as “Gene body methylation” (gbM) which typically occurs in the CG context. GbM occurs frequently in constitutively expressed genes whereas this is least observed for genes with highly variable expressions and therefore it has been hypothesized that the gbM may not modulate expression during development or response to the environment (Zilberman, 2017). Interestingly, although gbM primarily occurs on long evolutionary conserved genes, it is absent in the fungal genome. CRE methylation can change the accessibility of that element to TFs and thus activate or repress gene expression depending on the role of that CRE for the target gene and its accessibility profile upon methylation. For instance, loss of methylation in a short repeat that resides in the upstream promoter of the *FLOWERING WAGENINGEN (FWA)* gene in Arabidopsis has shown to make this element accessible for the transcription machineries which eventually increases the expression of *FWA* gene and delays the flowering of the plant (Soppe et al., 2000; Zhong et al., 2021).

The different epigenetic marks we have discussed above are dynamic in nature. Numerous epigenetic reprogramming events have been reported during the various developmental stages of plant’s life cycle and in response to various environmental stimuli including biotic and abiotic stresses (Zhang et al., 2018). The major purpose for reprogramming DNA methylation during gametogenesis and embryogenesis is to protect genome from TEs (Rajkumar, Gupta, Khemka, Garg, & Jain, 2020). As an example, a study in rice endosperm revealed the reprogramming of methylation in different contexts: non-CG methylation is reduced globally whereas CHH methylation of small TEs is increased in embryos, a pattern that is conserved among angiosperms (Zemach et al., 2010). There is a growing interest to understand the role of environmental stimuli on a plant’s epigenomic marks and whether the plant can preserve past environmental cues as a memory in epigenomic marks. Studies on natural variant of epialleles suggest that genetic differences in natural population have strong influence on the pattern of DNA methylation but it is uncertain whether this difference leads to adaptation of plants to its native environment (Dubin et al., 2015). Studies have demonstrated that plant DNA methylation can be altered at individual locus under both biotic and abiotic environmental stresses (Wang et al., 2011; Wang et al., 2014; Wang et al., 2020). One such example of DNA methylation for CREs can be taken from the recent study conducted by Wang et al. (Wang et al., 2020). The authors showed that in rice DNA methylation of a miniature inverted repeat transposable element (MITE) located in the promoter region of a key salt responsive gene *OsHKT1;5* increases under salinity stress which further recruits a methylation reader OsSUVH7 at MITE. A MYB transcription factor OsMYB106 along with a chaperon regulator OsBAG4 bind to this transcription complex and facilitate the expression of *OsHKT1;5*. *HKT1;5* has been well studied in plants and encodes a Na^+ selective transporter and helps to maintain the Na^+/K^+ homeostasis during salt stress (Kobayashi et al., 2017).

Now a days there are three primary methods available to detect DNA methylation; (1) bisulfite conversion and sequencing (BS-seq), (2) differential enzymatic cleavage of DNA, and (3) affinity capture of methylated DNA. One can detect DNA methylation either for some specific locus or search the marks genome-wide. For locus specific detection, first the target needs to be amplified and then the methylation can be detected either by enzymatic cleavage or BS-seq. A combination of methyl sensitive and insensitive restriction enzymes can be used to detect methylation for some specific locus of interest. Methylation sensitive enzymes can only cleave at unmethylated targets of the restriction site and this property has been exploited to detect DNA methylation for the given restriction site. Affinity capture of methylated DNA technique uses antibody immunoprecipitation method that utilizes a 5-methylcytidine antibody to specifically recognize methylated cytosines for enrichment and can further be sequenced using an NGS platform. In this section we will only discuss the bisulfite conversion method in detail for whole genome DNA methylation analysis. The general principle of bisulfite conversion is fairly simple: the treatment of genomic DNA with sodium bisulfite converts cytosine (C) residues to uracil (U) and leaves 5-methylcytosine residues unaffected. To construct libraries for whole genome BS-seq, genomic DNA usually undergoes sonication to obtain 100–500 bp fragments and then is ligated with appropriate adapters for latter sequencing steps. These small fragments are then subjected to sodium bisulfite conversion, few rounds of amplification and finally sequenced using high-throughput NGS platform to provide single base resolution of the DNA methylation call. One key consideration of BS-seq is the conversion rate of C to U which can be optimized on the small chloroplast genome. While calling for DNA methylation, any mismatch of T in query sequence aligned to C in the reference genome can be considered as unmethylated base. For a cytosine base of a given position in a reference genome we would expect multiple BS-seq reads to align and therefore a binomial test is usually required to call whether the base is methylated. Methylation profiles are usually analyzed for different contexts and are tested for a given window length of a genome.

In the following section, we have mentioned some examples of DNA methylation for abiotic stress responses. This raises the question: what is the scope for exploiting this methylation process to improve resilient crop production that can withstand climate change? For the past two decades scientists have started to understand the process of gene expression regulation by DNA methylation in the context of abiotic stresses. Unfortunately, we still have very little understanding of how these stress memories pass to the next generation of plants to maintain the regulatory process. More studies for the heritability of these stress memories may lead towards the development of stress tolerance crop breeding by improving and exploiting existing epigenetic engineering tools.

33.4.2 The role of histone modification for the regulation of gene expression

Histone modification is one of the few key epigenomic mechanisms that regulates gene expression in plants and plays a critical role in maintaining spatiotemporal transcriptional dynamics for a multicellular organism at various developmental stages and in response to different environmental stimuli. Typically, eukaryotic DNA is wrapped around histone proteins and undergoes several degrees of folding to form compact chromatin structure that limits the accessibility of DNA to its binding partners. These histone proteins can be modified posttranscriptionally and as a result can alter the accessibility of the associated DNA. Five major classes of histone exist in the eukaryotic system: H1/H5, H2A, H2B, H3 and H4. The latter four are usually referred to as core histone molecules whereas H1/H5 are known as linker histones (Park & Kim, 2020). The core histones have a conserved motif called the histone-fold domain which is a globular structure and has a dynamic histone-fold extension which is called “histone tail” (Zheng & Hayes, 2003). Two of each of these core histone proteins form a histone octamer which binds and wraps about 146 bp of DNA molecule (~1.7 turns of DNA helix) (Li, Carey, & Workman, 2007; Luger, Mäder, Richmond, Sargent, & Richmond, 1997). The linker histones additionally wrap around another 20 bp of DNA and connect to the next histone octamer which eventually lead to an array of histone octamers wrapping the DNA molecules. This primary extended nucleosome array structure undergoes secondary and tertiary folding and forms the compact chromatin structure (Caterino & Hayes, 2007).

The tails of the histone molecules that interact with DNA, and the globular domain of histone protein surface far from DNA molecules are both subjected to numerous different kinds of posttranslational modification which also modulates gene expression. These modifications include methylation of Lysine and Arginine; acetylation, ubiquitination, ADP-ribosylation, and sumoylation of Lysine and phosphorylation of Serine and Threonine. These modifications which are associated with transcriptional activation such as H3K4 methylation (read as methylation at the 4th Lysine residue of the H3 histone protein) are referred to as “euchromatin modification” whereas modifications that are localized for inactivation of gene expression such as H3K27 methylation (read as methylation at the 27th Lysine residue of the H3 histone protein) are referred to as “heterochromatin modification” (Caterino & Hayes, 2007). Another major class of chromatin remodeling involves ATP-dependent alteration of histone-DNA contacts which use ATP hydrolysis energy for this remodeling, can form DNA loops and slide the nucleosome position which eventually changes the accessibility of DNA to TFs. Another level of chromatin complexity may occur in the nucleosome due to the variants of core histone molecules. The recruitment of new histone variants can affect the existing posttranslational modification and alter nucleosome stability and the interaction between nucleosomes.

Chromatin immunoprecipitation (ChIP) for modification-specific antibody such as antibody targeted for H3K9Me (read as methylation at the 9th Lysine residue of the H3 histone protein) coupled with microarray or DNA sequencing is a widely applied method for genome-wide mapping of DNA–protein interactions (Tao, Feng, Zhao, & Guan, 2020). Unfortunately, this method provides high background signals and suffers for a higher false positive rate. Several other alternative methods have been developed in past few years such as DNase I hypersensitive sites sequencing (DNase-seq) (Crawford et al., 2006), Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIRE-seq) (Giresi, Kim, McDaniell, Iyer, & Lieb, 2007), Micrococcal Nuclease digestion with deep sequencing (MNase-seq; Kent, Adams, Moorhouse, & Paszkiewicz, 2011) and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013) which aims to profile the accessible chromatin regions (ACRs), not the specific type of modification that causes the activation or repression. All of these techniques exploit the property of open chromatin structure but differs on whether the open-nucleosome free region or closed nucleosome-covered regions of DNA is subjected to NGS-sequencing. In this section we will only discuss the principle and methodology of ATAC-seq technique which became a popular choice for its simplicity and robustness. This method uses a modified Tn5 transposase with sequencing adapters that integrates itself directly into the open nucleosome-free region of DNA and initiates the first step of the sequencing library construction. Insertion of Tn5 leaves the footprint of chromatin regions that are accessible which are then sequenced by NGS platform after amplification. However, for plants, a major source of contamination of organelle genomes can reduce the representation of nuclear genome. Therefore a new

modified protocol has been developed which coupled fluorescence-activated nuclei sorting (FANS) prior to ATAC-seq initiation (Lu, Hofmeister, Vollmers, DuBois, & Schmitz, 2017). In order to identify ACRs from ATAC-seq data, one would expect to observe enrichment of read alignments around the open chromatin region compared to the region that are covered by nucleosomes. Therefore commonly used peak calling tools which identify ACRs such as MACS2 (Zhang et al., 2008) or HOMMER (Heinz et al., 2010) test for enrichment of read counts for a given window of the genome compared to various length of random genomic backgrounds. These tools were primarily developed for ChIP-seq data and later have been modified to adapt for ATAC-seq data analysis. Currently, HMMRATAC (Tarbell & Liu, 2019) is the only peak calling tool which exclusively has been designed for ATAC-seq data but has not been implemented in many studies yet.

Identification of ACRs circumvent the unique difficulties in identifying CREs genome-wide by reducing the search space from the whole genome to few thousand peaks. ACRs also provide the opportunity to study the cis-regulation of gene expression and how GRN can be modulated for different developmental stages of plants or tissue type or by the external environmental stimuli. For instance, a study for open chromatin landscape in Arabidopsis for root hair and non-hair cell types revealed a root hair cell transcriptional regulatory module which is driven by ABA INSENSITIVE5 (ABI5) and MYB33 TFs (Maher et al., 2018). Another study integrated time-series transcriptome data and the patterns of nucleosome-free chromatin to reveal the environmental gene regulatory influence networks which regulate the gene expression in response to high temperatures, water deficit, and agricultural field conditions in five natural accessions of tropical Asian rice (Wilkins et al., 2016). Another recent open chromatin landscape study in sorghum identified drought-induced regulatory module and the variation in core drought-inducible signatures that associated with plant's water use efficiency (Parvathaneni, Kumar, Braud, & Eveland, 2020).

One may ask the question how conserved these CREs and the GRNs are across the long branches of the evolutionary tree and can the knowledge of one plant species will be helpful for another species in the context of similar stress conditions. Several studies have identified interspecies conserved noncoding sequences (CNSs) for plants (Burgess & Freeling, 2014; Turco, Schnable, Pedersen, & Freeling, 2013; Van de Velde, Van Bel, Vanechoutte, & Vandepoele, 2016). Lu et al. (2019) recently showed that such CNSs are enriched in distal ACRs that are very far from target genes. Therefore the understanding of these intraspecies conserved stress GRNs regulations offers an in-depth picture of the stress GRNs and opens the scope for manipulations of CREs/TREs for stress tolerant crop improvement.

33.5 Protein structure provides vital information of function during salt stress

Transport proteins are fundamental to life because they coordinate the movement and distribution of solutes between different parts of the plant. Most of these ion transporters are responsible for regulating the metabolism, salt tolerance or sensitivity and development of plants (Zhao, Zhang, Song, Zhu, & Shabala, 2020). In some cases, the expression pattern of a particular transport protein may not differ between a tolerant or sensitive genotype under say, salt stress (Shohan, Sinha, Nabila, Dastidar, & Seraj, 2019). In such cases, it is difficult to understand how the plant defends itself in order to survive and thrive under salt stress. Under these circumstances, integration of structural and functional studies can provide insight into the mechanism of salt stress tolerance. Vital information gathered by matching structural information among homologous transporters and their quantitative expression data have allowed the prediction and testing of key structural elements. Specific amino acid residues were found to be implicated, which might help in conferring salt tolerance through substrate specificity/selectivity and/or protein-protein interaction at the molecular level. For such transport proteins, discovery of the substituted amino acids, prediction and validation of the altered molecular structure led to an understanding of their differential function in tolerant and sensitive genotypes. In this section, we illustrate the power of combining structural modeling and functional assays in order to understand how transport function can enhance the ability of plants to fight and survive salt stress.

33.5.1 Variation in protein structure contributing to salinity tolerance

Salt exclusion is of vital importance to help alleviate the effect of salt stress in any crop. High soil salinity leads to accumulation of toxic levels of Na^+ ions which interfere with photosynthesis, halt nutrition, and stall the development of the plant. As such, excess Na^+ needs to be kept away from cellular cytoplasm, especially in the shoot (reviewed in Deinlein et al., 2014). Several studies have shown how structure-function relationships play a crucial role in a more effective Na^+ exclusion of the plant and better tolerance to salt stress. High Affinity K^+ Transporters (HKT), in particular HKT1;5, which are associated with the root-shoot vasculature have been shown to minimize the accumulation of Na^+ in shoot tissues (Horie, Hauser, & Schroeder, 2009). The crystallization of the HKT1;5 protein has not been

possible up to now. Therefore high quality modeling has been done using the bacterial K^+ transport members of the Ktr/Trk subfamily as templates. In one study it was shown that the HKT1;5-A transporter from *Triticum monococcum* (TmHKT1;5-A) and HKT1;5-D from *Triticum aestivum* (TaHKT1;5-D) selectively conduct Na^+ ions at different affinities and rates (Xu et al., 2018). This study found 27 differences in amino acid residues between TmHKT1;5-A and TaHKT1;5-D. Among these residues, substitution of six residues were predicted to cause sufficient structural change of the transporter proteins and impact their transport capacity and function. These changes from basic to acidic residue or positive to negative charge significantly impacted the local structure.

Another study conducted on different rice varieties showed that HKT1;5 has functional variability among salt-sensitive and salt-tolerant varieties including in the halophyte *Oryza coarctata* (or *Porteresia coarctata*) (Shohan et al., 2019). The major variation was found to be due to four altered amino acids. In this study, molecular dynamics simulation modeling showed the positioning of the substituted Aspartate and Valine on opposite ends of the membrane in the tolerant varieties. The dynamics showed that the substituted Valine (which is smaller than the Leucine present in sensitive genotypes) was unable to generate a strong hydrophobic network, resulting in alteration in pore rigidity and easy transport of Na^+ away from the shoot. On the other hand, the presence of the substituted Aspartate in the shoot to root interface created frequent polar interactions in the extracellular loop permitting less constriction at the pore and easy efflux of Na^+ compared to the Histidine at the same position in sensitive genotypes (Shohan et al., 2019). Homology modeling and simulation study of HKT1;5 protein from the wild halophytic relative *O. coarctata* suggested that the relatively lower Na^+ affinity of this transporter was due to four key amino acid changes in the loops on the extracellular side (E239K, G207R, G214R, L363V) (Somasundaram et al., 2020). This was validated by Na^+ transport assays after reciprocal site-directed mutations. Another study conducted with HKT1;2 from the halophyte *Thellungiella salsuginea* (TsHKT1;2) found that this protein transports K^+ in the presence of Na^+ in yeast (Ali et al., 2016). TsHKT1;2 and most other HKT1;5 sequences have aspartate in the second pore domain, whereas in all other cases, the presence of asparagine was reported. Mutation studies have shown that replacing asparagine with aspartate in HKT1 type transporters leads to altered cation selectivity and uptake dynamics (Ali et al., 2016).

Conversely, one study conducted on rice HKT1;3 (OsHKT1;3) showed that a change in amino acid may not have any effect on salt stress tolerance at all. Five SNPs were found in the coding region and among them four were synonymous substitution (A798C, G2083A, T2101C, C2122T) and only one was nonsynonymous (C3598G) which changed the amino acid at position 200 from Leucine to Valine. But the position of the amino acid was in the third transmembrane segment of the OsHKT1;3 protein and the authors asserted that there was no effect on its transport capacity due to this change (Do, Hoang, Le, Tang, & Nguyen, 2018).

Transporters like HKT1;5 seem to play a central role in adaptation to salt stress and despite the importance of their structure and resultant function as discussed above, there may be an added layer of regulation. It has been recently shown that the promoter of rice HKT1;5 is subject to epigenetic control during salt stress. A transcriptional complex containing a methylation reader of a transposable element, a Myb binding site with OsMyb106 and a bridging protein was shown to regulate expression of HKT1;5 in rice (Wang et al., 2020).

Boron is a solid metalloid which is passively up taken by the plant root from soil. Although high soil boron toxicity is widespread worldwide, boron tolerant plants accumulate lower concentration of boron compared the sensitive ones. The main mechanism for boron toxicity tolerance is related to limited entry of boron in the form of boric acid (BA) and removal of surplus BA through leaves (Princi et al., 2016). In barley, two particular genes namely *NIP2* and *Bot1* are mainly responsible for maintaining boron concentrations. Both of these genes encode proteins with transmembrane alpha-helices and reside on the epidermal root cell. *Bot1* plays a central role for imparting tolerance to plants growing in soils with a high content of boron through efflux of borate from cells (Princi et al., 2016). One study defined the function of *Bot1* using cell-free synthesis through combination of molecular dynamics simulation, site-directed mutagenesis and nanotechnology (Nagarajan et al., 2016). The authors found that the variant sites L234H and T541M occur in the barley cultivar Haruna Nijo where H234 resides in the interhelical loop, adjacent to the fully conserved Arg-235 facilitating transport function. On the other hand, Met-541 resides in the intracellular loop in the region of low conservation. Mutagenesis study showed that L234H in *Bot1* is critical for the function of the protein while T541M has no effect. The other changes found in *Bot1* alleles from the cultivars Tadmor, Alexis, and WI4304 (N108S, K183E, and Y195F) do not impact the functional atomistic model as are located in the noncritical region. This study also showed that *Bot1* has a Na^+ ion binding site which is essential for its conductance (Nagarajan et al., 2016).

The Dehydration-responsive element-binding (DREB) transcription factor is an important regulatory molecule involved in stress signal transduction pathways, hormone response, and plant derived development functions. Members of the DREB family carry a conserved DNA binding domain known as EREBP/APT2. Analysis of amino acid sequences of DREB proteins identified the presence of a conserved nuclear localization signal and another conserved

serine/threonine rich region adjacent to the EREBP/AP2 domain (reviewed in Agarwal, Agarwal, Reddy, & Sopory, 2006). Several residues important for DNA binding activity of EREBP/AP2-type proteins have also been identified, for instance arginine and tryptophan. Other key residues include a glycine within the AP2 domain that forms hydrogen bond with an alanine (Allen, Yamasaki, Ohme-Takagi, Tateno, & Suzuki, 1998). Mutations in this glycine residue have been demonstrated to impair the function of the Arabidopsis APETELA2 protein, possibly due to changes in the structure (Jofuku, Den Boer, Van Montagu, & Okamuro, 1994). Comparison of DREB1 proteins in multiple durum wheat (*Triticum turgidum*) varieties have also revealed a SNP within AP2 domain in the highly salt tolerant genotype (Mondini, Nachit, Porceddu, & Pagnotta, 2012).

The WRKY family of transcription factors play critical role as transcriptional regulators by repressing and/or activating different plant processes, including abiotic stress responses (Li, Pang, Lu, & Jin, 2020; Rushton, Somssich, Ringler, & Shen, 2010). Several members of the WRKY family have been implicated in drought and salt stress tolerance in multiple plant species (Mondini et al., 2012; Shi et al., 2014). Comparative analysis of *WRKY1* transcripts in durum wheat varieties revealed two SNPs associated with salt-tolerant genotypes (Mondini et al., 2012). The first SNP consists of a A/T transversion in the moderately salt-tolerant genotype Cham1. The second SNP consisting of a G/C transversion was found in the highly salt-tolerant genotype J. Khetifa. Both of these SNPs were located near the WRKY domain and resulted in amino acid substitutions. Thus these SNPs could potentially affect the DNA-binding function of the WRKY domain and possibly contribute to increased salt tolerance in the aforementioned genotypes.

The *Salt Overly Sensitive 2 (SOS2)* gene is essential for homeostasis of Na⁺ and K⁺ level in plants. *SOS2* codes for a serine threonine type protein kinase with an N-terminal catalytic domain which is also seen in yeast SNF1 kinase (Halfter, Ishitani, & Zhu, 2000). Analysis of the *sos2* mutant allele sequences showed that both the N-terminal catalytic domain and C-terminal regulatory domains are essential for the *SOS2* to function properly. Another study in *Arabidopsis thaliana* showed that changing glycine to glutamate in the recessive *sos2-5* allele abolishes the *SOS2* autophosphorylation. Therefore it is evident that this mutation causes change in the catalytic domain of the *SOS1* protein structure, making the plant susceptible to salt stress (Liu, Ishitani, Halfter, Kim, & Zhu, 2000).

33.5.2 Future prospect in substitution-mediated enhanced salt tolerance

Knowledge of structure and transport function will help us to understand the impact of substitution events on permeation, particularly by analyzing the atomic structure of 3D models of transporters. Coupling this information with molecular dynamic simulation can empower protein engineering for better transporter function and improved salt tolerance. With the advent of CRISPR-Cas technologies, such engineering is becoming more feasible. The knowledge of structural variation with substitution of amino acids will also help in the accurate annotation of genes and provide better understanding of the significance of variation of homologous and divergent genes in an evolutionary context as well.

33.6 High performance computing in comparative genomics

Modern science is data driven. With the advancement of digital sensors, we are generating more and more data. As explained by the Wired magazine in its July 2018 cover—“The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data” (<https://www.wired.com/>). This states how the science is shifting. In fact, a lot of our understanding in every discipline is generated by analyzing a massive amount of data. For example, the Rubin Observatory (also known as LSST) located on a mountaintop of Chile is equipped with a 3.2 gigapixel CCD that can generate 1.28 petabytes of images per year for astronomical surveys. The Large Hadron Collider (LHC) built by the European Organization for Nuclear Research (CERN) produced approximately 25 petabytes of data every year by 2012. Commercial giants like Google, Yahoo, Microsoft and Amazon are also gathering information for prediction and effective extension of their marketing policies. These data are also being used for economic and social science research. Biological science is very understandably also in this race. Advancement in sequencing technologies has led to an enormous amount of data being deposited into databases. The 1000 genome project, the 3k rice genome project, etc. are few examples of large initiatives where large amount of data is being produced. Sequences of more and more crops and their variant accessions have also been deposited and there is a need to relate the SNPs with important traits for crop security in the face of climate change. GWAS not only of human diseases but also for crops with important trait variations need immense computational power. Generating and managing large amount of data brings with it new challenges but more so because the data needs to be understood and made meaningful. Development of proper tools for comparative genome analysis, structural and functional annotation of novel genes and proteins, understanding regulations of genes, finding related metabolic pathways and discovering epigenetic regulation have become crucial. Such information

is also invaluable for pharmaceutical industries, who need information on *in silico* target identification for drug discovery. All these boils down to the need for acquiring and managing computational requirements as well as skills. In general, we refer to supercomputers for addressing our computational needs for analyzing big data. But a supercomputer is expensive to build and is beyond reach of most researchers. The best way to resolve the problem is to break down big data into usable information, understand the components of computational needs for analyzing biological data, and make the best use of the resources we have.

The basic requirement of analyzing big data in biological science, generally referred to as bioinformatics, is computational power. However, the hardware requirement varies according to the type of analysis, type of tools being used, amount of data under processing, etc. For example, assembly of a bacterial genome using the Unicycler tool on a standard workstation with CORE i7 processor and 16 GB random access memory (RAM) takes about 2–3 h. But assembly of a large genome like rice (~400 Million bp) or human (~3 billion bp) in it will be nearly impossible. Even in the example for microbial genome assembly, the required time will vary depending on arguments used for analysis, type of hardware or operating system. This problem is not unique to biological science only, rather it is universal. The computing community consortium recommended guideline to prepare or manage the system involved for such mammoth tasks (Bryant, Katz, & Lazowska, 2008). For managing our requirements, we can adopt the right system. According to the guideline, the following factors must be considered.

Storage: The data can be from two sources: from the sequencing machine itself, or it can be downloaded from public databases. Whichever the source is, a good practice would be to store the data in a central location. This is usually a computer with a large storage facility and is connected to network through which the data can be accessed by the computer where the analysis is taking place. It will eliminate the overuse of storage due to multiple copies downloaded by multiple users. However, local storage (the computer where the analysis is taking place) also has to be in consideration. Because, most of the analysis tools generate temporary files while performing the task and they can be quite large in size. Advancement of magnetic disk technology (technology behind traditional hard disk) has made the storage devices cheaper. However, it has a rather slow read/write speed compared to the processing power available, which often becomes the rate limiting step in overall performance. A traditional hard disk offers about 125 MB/s data transfer (read/write) speed. If equipped on a 32 core processor-containing computer, the data is processed at a faster rate than the reading of the information from HDD or placing them back. Recently developed SSD technology which offers 10x to 20x faster transfer rate has eliminated this problem. But they are rather expensive. One could balance this by choosing a SSD storage device as primary where the operating system is installed and also being as temporary storage while the analysis is ongoing. Whereas, magnetic disk technology storage can be used as secondary, where the final output will be stored. With this the initial reading of the data and final writing of data (which is on a traditional HDD) will be slower but will provide faster processing.

Network connection: Let's say we are storing the data on a central computer storage which is dedicated for this purpose only. The analysis will be done on a different high-performance computer. To minimize the delay to access the data, a good network connection will be required between these two facilities. To scale this up, the data stored at the central facility, can be shared among researchers across the globe. In this case a high-speed internet connection will be required.

Processor: The processor power is usually measured by number of cores, the operational unit of a processor. The popular laptop or desktop processors generally contain 4–8 cores and are not suitable for this task. Usually the processor core count is a major factor. The more the number of cores, the higher the number of processes that can be handled at a given time, which in turn speeds up analysis. A 32-core processor is a good starting point. However, it is highly dependent on the task in hand. Assembly of a single rice genome for example, will take 24–30 h on a 32-core computer. But a task like 3k rice genome, we may have to multiply the power to several folds and also use parallel processing on multiple clusters. Tasks like finding splice variants or epigenetic landmarks, etc., are less processor intensive than assembly, but the number of samples in the study may need to be considered.

RAM: RAM acts as temporary memory while the process is running. The larger the file size being analyzed, the more RAM it will require to process. For example, assembly of a bacterial genome of around 3 Mb, a computer with 16 GB RAM should be sufficient. Whereas, rice genome is about 400 Mb and therefore the sequence file to cover this length will be significantly larger. It would take 128 GB or more RAM to operate smoothly.

Operating system: One can choose any of the operating systems they are comfortable with. However, the availability of tools/software in use will be the determinant. A wider community adopted the Linux environment to develop most of the tools. Although some of the tools also have Mac and Windows versions also, the option for different types of analyses becomes limited in non-Linux environments because of the lack of compatible tools. Switching between operating systems is also not convenient. Moreover, Linux offers better multithread management over others. For example, we

want to launch a GATK analysis for SNP determination in a bacterial genome sequence in a computer containing 32 core processor. This analysis does not need to engage all the computational resources. For example, users can allocate 24 cores for the GATK analysis and use the rest for another task in a Linux environment. Moreover, using multiple terminals we can do the same task more efficiently. For example, in four separate terminals, we can start assembly of four separate bacterial genomes where we can assign 8 cores for each of the processes. The result will take almost the same time to finish a single process as assigning 32 cores for each, but in the latter case 4 genomes can be analyzed for the given time. But for a large genome like rice, it will take larger amount of resources per process. Another question that generally arises is: which version of the operating system should be chosen—the server or workstation version? Unless it's a requirement from the tools in use (like bigsdb, which requires local database management system), there is not much difference between the two versions. A server version will provide better access control over storage, and tools/software in use, especially if it is accessed remotely. Especially for centrally maintained data storage, one may want to limit the access of data to a specific number of users and not to all. A server edition of an operating system will provide better management of such access control.

Graphics: Most of the tools for assembly, annotation, alignment, etc., do not need additional graphical processing. However, analyzing the results and presenting them graphically, one may want to use software like “R” which is very popular among data analysts in recent days and will require a lot of graphics processing power. The simulation-based and molecular dynamics study tools would require a decent amount of graphical processing power. For analyses involving these tools, a good graphics processor (graphics card) could be useful.

Both processor and RAM requirement are dependent on the tools in use. For example, an assembly tool is more processor intensive, where as a multiple alignment tool is more RAM intensive. Again, a big dataset like rice genome will require larger amount of RAM for processing compared to a bacterial genome. Similarly, multithread enabled tools can perform a job faster with higher number of processing units (core) since they can efficiently split the task to multiple cores and return the output in significantly less amount of time. Therefore proper hardware needs to be chosen depending on the intended use. This brings us to the option of cluster computing. The concept of cluster computing is to split the job in several processes and reassemble them to the final output. The computing units can be several computers connected as node and task is distributed through a job manager (Fig. 33.2). Essentially this follows the principle of building-up of a super computer. A supercomputer consists of 100s or even 1000s of computers as nodes, where each node is powerful enough (32 core processor and 128 GB RAM for example) to perform a task in reasonable amount of time. If one has to start small and plan to build up later on this, building up such an architecture for a cluster computer may be a good start.

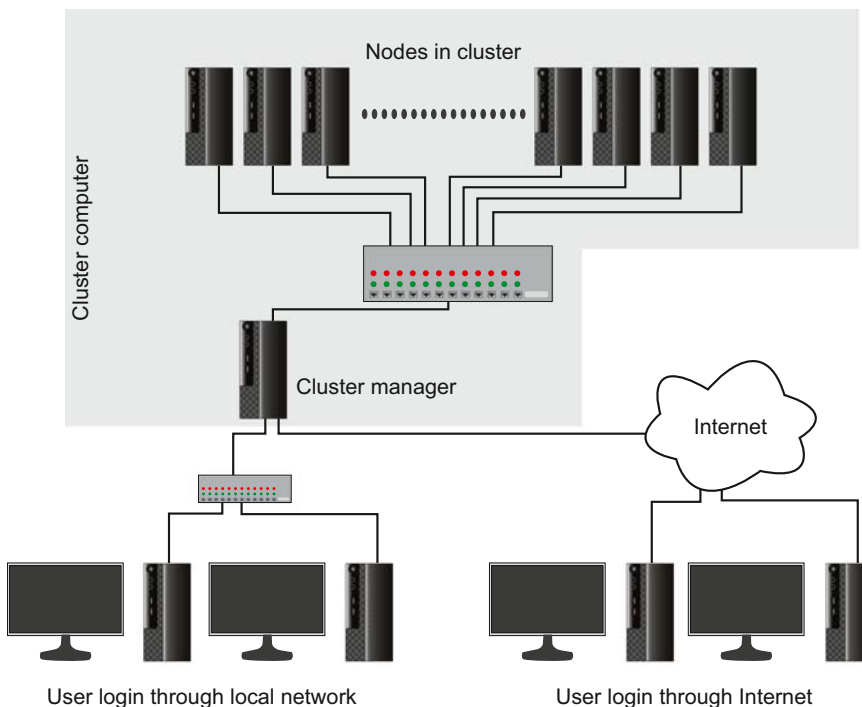


FIGURE 33.2 Cluster computing. This is a process of compiling several computer powers to represent a more powerful computer. A task is distributed among each of the computers known as node. The distribution and reassembly is managed through a managing system. This system can be accessed through a remote computer connected under local network or through the internet.

Another creative use of this cluster computing concept is employed by CERN to process the LHC data. This is employed by a software called Berkeley open infrastructure for network computing (BOINC) (Fig. 33.3). This software is connected to the data server and manages connection to hosts all over the world. Here a host can be anyone who is willing to share the ideal resources on their personal devices with the client software installed in them. The BOINC distributes the data to the client and reassembles the feedback.

In addition to the local computational unit or a supercomputer, a new emerging choice is cloud computing. The commercial giants like Google and Amazon built storage facilities and computing power that can be used with a subscription fee. The user only has to build the set of instructions for the computational units to process the data. The process is nothing different from the command lines prepared for processing in a personal computer, but are prepared for the cloud server. Necessary adjustments specific for the particular cloud servers are generally provided by the service itself.

Efficient uses of the computational units: Now that we have a general idea about the computational units, we will discuss the tools for the processes to plan a task to efficiently use the resources available. Most common uses of NGS data are assembly of the short reads into genome/transcriptome or find the variants by comparing to a reference genome. Assemblies are of two types: *de novo* and reference guided. The *de novo* assembly process consists of several steps described in Fig. 33.4. Depending on the assembly program in use, the command to the process can be automated in a single step or multiple steps where user has to initiate each step. The latter may seem laborious but gives more control to fine tune the process. In *de novo* assembly, the process varies depending on the tools being used but the basics are pretty similar. First, we may want to check the quality of the reads and remove the bad reads. Next step is to assemble the reads into a contig. This step takes a lot of processing power. Therefore if we assign maximum available resources into this, it will take minimum time to complete. Most of the recent releases of the assembly programs are multithread-enabled, but the default parameters are set to a lower number (usually 4) to allow universal use. Next steps are aligning the contigs, and reassembling them into draft genome. These steps require a lot of RAM but considerably lower processing. For assembly of a large genome like rice or human genome, we may want to deploy all the resources available. But for small genome like virus or bacteria, assembly takes a lot less resources and assigning all of them is not necessary.

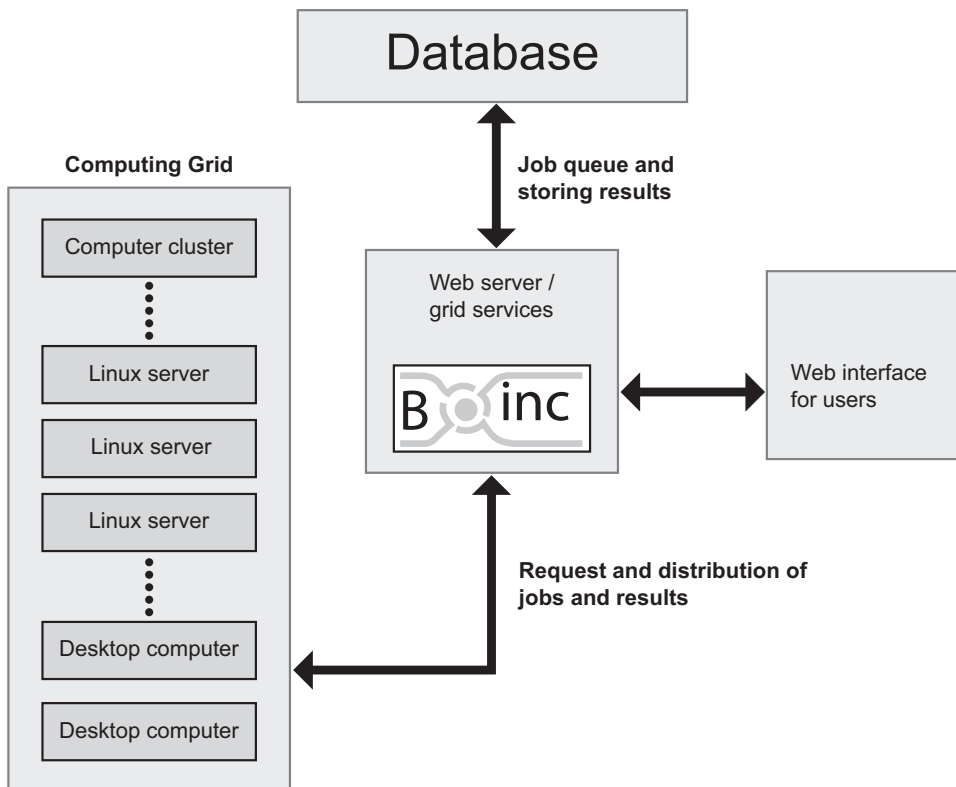


FIGURE 33.3 BOINC manager. The BOINC web server manages job distribution along the grid which can be computer hosts inside the network or outside hosts willing to share the ideal resources. The server collects the results and assembles into the database. Figure is adapted from <https://boinc.berkeley.edu/>.

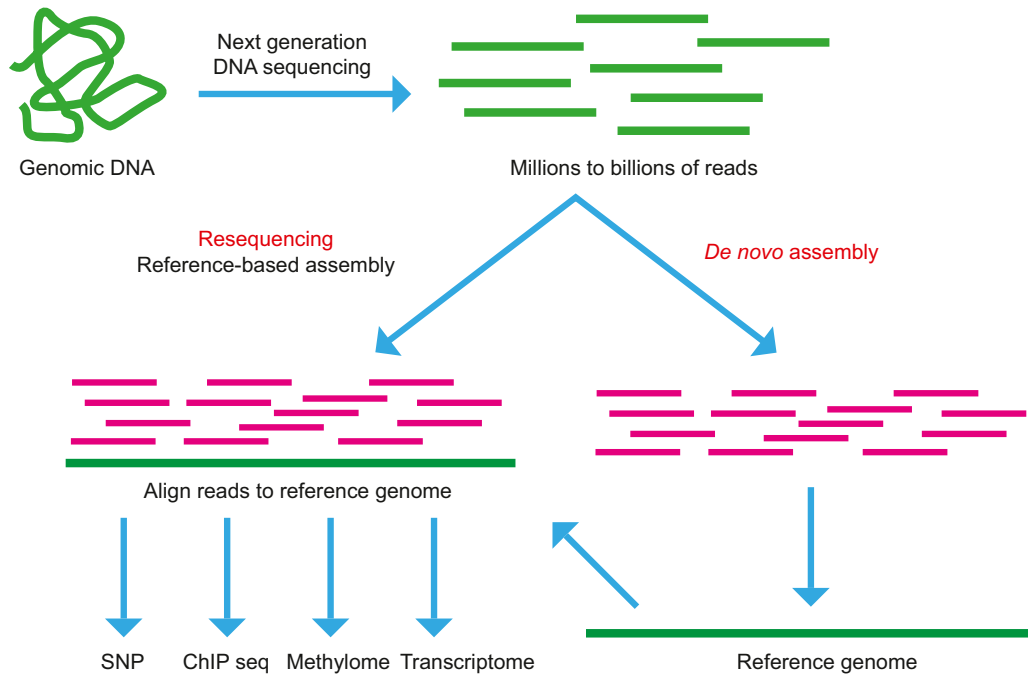


FIGURE 33.4 NGS data generation and processing. Genomic DNA (or RNA) is processed according to experiment. Sequencing machine generates millions to billions of short reads. These reads are processed according to experiment design. Short reads are assembled with or without a reference guide. A *de novo* assembly will generate contigs which can be realigned and curated to produce a reference genome. In a reference-guided assembly, reads are compared to reference genome and deviations from reference are determined. Based on experiment design, these data can be further processed for SNP, methylome, ChIP Seq interpretation. *NGS*, Next generation sequencing; *SNP*, single nucleotide polymorphism.

For reference-based assembly, the short reads are aligned to a reference genome. The process is similar in variant analysis, as well as ChIPseq and methylome analysis. The latter however requires a lot of RAM. The process automatically will use maximum available RAM. Therefore when we plan these studies, it is best to start with sufficient RAM. We often use one or more segments of genome under these studies. Defining these regions to the alignment tool will greatly improve the performance since it has to align only to the target regions rather than the entire genome.

Programming language: A general myth among biologists is that bioinformatics requires a lot of computer programming and therefore learning programming language is a must. Bioinformatics can be divided into two branches: developing the tools for a specific need and use of the tools for actual analysis. The earlier task requires a lot of programming knowledge and is usually done by someone with computer science/software background. In this case, a specific problem is presented to them. For example, assembly of NGS data, where the task is to align millions of 100-bp or similar reads into the final contig. A programmer will find out a smart way to place the correct sequence one after another or over one another in case of duplicates. Most of the cases, they do not need to know what the sequence means or its significance. They just need to be able to read them and place them in order. This is usually done with programming languages like C++ or python. On the other hand, biologists are the user of the tools. They will use the tools developed by the earlier group and apply it on real dataset and experimentally prove if the result is correct and produce a meaningful contig that represents the given organism. However, for efficient operation and efficient use of the resources, a little understanding of programming language is helpful for the latter group. For example, the assembly process in multiple steps requires human input in each step. We have to wait one process to finish before we start another one to start. This is laborious and becomes complicated when we want to use parallel processing. Writing a small script (a set of commands placed together which will provide instructions one after another and will respond accordingly) can easily handle this task automatically. Often, the output of a tool does not directly fit as input of the next tool. Manual modification/conversion is a mammoth task. For example, the local blast output would provide a multiple column data. For each sequence there would be multiple line results. But we may be processing blast results of 100 sequences and from each result, we want to pick only the best hit for those 100 sequences. Usual process would be to find the best hit value for each of the sequence and copy the information to a new file then cut out the accession and sequence for them. This may take several hours of intensive work. But with a few lines of code, we can do the same within a few seconds.

A programming language such as R can analyze the results and present them most efficiently in graphical format. As Jeffrey Perkel, the technology editor at the *Nature Journal*, aptly described (Perkel, 2021)—a little understanding in Unix/Linux can be helpful to access a file without even opening it and search for specific information and used to advantage for many analyses.

In recent days, more and more scientists are becoming interested in programming languages. Programming languages such as R, Python, and Perl, are extensively being used by biologists for data processing analysis and graphical representation. All these languages are equally capable for the analysis purpose. Most importantly, a wider community is using these languages and hence, new modules are constantly being developed for all of them, which we can utilize to process and represent our data more easily and efficiently with ease.

Computational skills: As mentioned earlier, developing a software/tool will require a lot of programming language and associated skills. Users (e.g., biologists) of these tools will not require such high-level programming skills, but a decent understanding will help a lot in processing the data. For example, while using a cluster computer, one might expect it to need a good amount of computer knowledge. But in fact, the architecture is designed usually by a network engineer and is usually operated through a management software. The user just has to place the usual commands only in a specific format for the management software to process. Skillful use of “awk” and “bash” in Linux is very helpful in this regard. For operating from a remote location (a computer different from the data stored or analysis is being performed), understanding on network architecture is helpful. A third-party software like BOINC also uses the same instructions but in a specific manner. To wrap up the requirements for biologists, it is the same for all, just will need some adjustment according to the system. For example, the command “bwa” will launch bwa alignment tool in a Linux system. But by using “ssh” in terminal we can log into a different computer, a super computer on a remote location for example, and “bwa” command there will launch the same program in that remote computer.

In summary, bioinformatics is the new era of modern biology. We may have limited resources to adopt the new section of science. But better understanding of the process will allow us to utilize the resources more efficiently and to explore the information that is already available. This will help us to find sustainable solutions for challenges like abiotic stress.

33.7 Conclusion

In summary, ensuring food security for the future is an uphill task, aggravated by global change in weather patterns and the need to feed an ever-increasing human population. Another all-important constraint for ensuring crop production are the dwindling reserves of fresh water. In addition, clearing of forests for logging and human habitation has drastically reduced plant diversity. The crops of the future need to be resilient to multiple stresses. Therefore it is imperative that we identify and study existing world-wide germplasm banks for plants with desirable traits such as abiotic stress tolerance. The DNA sequence of many reference genomes of cereal and legumes crops as well as their resequenced variant genomes have already been deposited in genome specific or plant genome databases as discussed in this chapter and GWAS as well as QTLs for important traits have already been published. Methods are also in place to generate large scale mutations for producing allelic variations for improvement of target traits using genome editing tools, including epi-alleles (Herbert et al., 2020; Zhang, Malzahn, Sretenovic, & Qi, 2019). The success in producing hardy crops which are commercially viable needs to gather speed. Bioinformatics tools have gone a long way in helping us handle and make sense of the enormous data already deposited. Further development of bioinformatics tools to understand the epigenome, RNA expression QTLs as well as small RNA expression QTLs will help in elucidating the fine-tuned regulation of stress responses in plants. Armed with this knowledge, modern revolutionary tools like CRISPR-Cas9 genome editing may pave the path for speedy crop improvement for the future. CRISPR-Cas9 has been used for gene knockout, which can be useful if we target susceptibility genes, or genes used by pathogens for establishment, or regulatory genes which lower gene expression under stress. Precise gene-editing of single and multiple genes is also possible by homology-dependent repair. Transcriptional control is also possible by gene editing of regulatory elements. Dead Cas nucleases can be used to recruit regulators to the promoter region, including activators, repressors, DNA methyltransferase, demethylase and so on, reviewed in (Zhang et al., 2019). Recently RNA viral transfection methods have been developed, where a single generation of plants are targeted for alteration of regulatory circuits which can enhance agronomic traits (Torti et al., 2021). Gene editing tools like CRISPR-Cas9 may be used in such a way that the superior plants produced are non-GMO and therefore can bypass time-consuming regulatory requirements, making it the one of the handiest tools for rapid crop improvement.

References

- Addo-Quaye, C., Miller, W., & Axtell, M. J. (2009). CleaveLand: A pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics (Oxford, England)*, *25*(1), 130–131.
- Agarwal, P. K., Agarwal, P., Reddy, M. K., & Sopory, S. K. (2006). Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. *Plant Cell Reports*, *25*(12), 1263–1274.
- Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R. R., et al. (2014). SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Research*, *43*(D1), D1023–D7.
- Ali, A., Raddatz, N., Aman, R., Kim, S., Park, H. C., Jan, M., et al. (2016). A single amino-acid substitution in the sodium transporter HKT1 associated with plant salt tolerance. *Plant Physiology*, *171*(3), 2112–2126.
- Allen, M. D., Yamasaki, K., Ohme-Takagi, M., Tateno, M., & Suzuki, M. (1998). A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *The EMBO Journal*, *17*(18), 5484–5496.
- Alonso-Simón, A., Kristensen, J. B., Øbro, J., Felby, C., Willats, W. G., & Jørgensen, H. (2010). High-throughput microarray profiling of cell wall polymers during hydrothermal pre-treatment of wheat straw. *Biotechnology and Bioengineering*, *105*(3), 509–514.
- Annacondia, M. L., Magerøy, M. H., & Martinez, G. (2018). Stress response regulation by epigenetic mechanisms: Changing of the guards. *Physiologia Plantarum*, *162*(2), 239–250.
- Axtell, M. J. (2013). Classification and comparison of Small RNAs from plants. *Annual Review of Plant Biology*, *64*(1), 137–159.
- Axtell, M. J. (2013). ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA (New York, N.Y.)*, *19*(6), 740–751.
- Axtell, M. J., & Meyers, B. C. (2018). Revisiting criteria for plant microRNA annotation in the era of big data. *The Plant Cell*, *30*(2), 272–284.
- Baker, R. L., Leong, W. F., Brock, M. T., Rubin, M. J., Markelz, R. C., Welch, S., et al. (2019). Integrating transcriptomic network reconstruction and eQTL analyses reveals mechanistic connections between genomic architecture and *Brassica rapa* development. *PLoS Genetics*, *15*(9), e1008367.
- Barciszewska-Pacak, M., Milanowska, K., Knop, K., Bielewicz, D., Nuc, P., Plewka, P., et al. (2015). Arabidopsis microRNA expression regulation in a wide range of abiotic stress responses. *Frontiers in Plant Science*, *6*, 410.
- Bhattacharjee, S. (2012). The language of reactive oxygen species signaling in plants. *Journal of Botany*, *2012*.
- Biscarini, F., Cozzi, P., Casella, L., Riccardi, P., Vattari, A., Orasen, G., et al. (2016). Genome-wide association study for traits related to plant and grain morphology, and root architecture in temperate rice accessions. *PLoS One*, *11*(5), e0155425.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120.
- Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., & Zhu, J.-K. (2005). Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, *123*(7), 1279–1291.
- Broman, K. W., Wu, H., Sen, S., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics (Oxford, England)*, *19*(7), 889–890.
- Bryant R., Katz R.H., Lazowska E.D. (2008). *Big-data computing: Creating revolutionary breakthroughs in commerce, science and society*.
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213.
- Burgess, D., & Freeling, M. (2014). The most deeply conserved noncoding sequences in plants serve similar functions to those in vertebrates despite large differences in evolutionary rates. *The Plant Cell*, *26*(3), 946–961.
- Caterino, T. L., & Hayes, J. J. (2007). Chromatin structure depends on what's in the nucleosome's pocket. *Nature Structural & Molecular Biology*, *14*(11), 1056–1058.
- Chen, L.-T., Luo, M., Wang, Y.-Y., & Wu, K. (2010). Involvement of Arabidopsis histone deacetylase HDA6 in ABA and salt stress response. *Journal of Experimental Botany*, *61*(12), 3345–3353.
- Chen, S.-Y., Su, M.-H., Kremling, K. A., Lepak, N. K., Romay, M. C., Sun, Q., et al. (2020). Identification of miRNA-eQTLs in maize mature leaf by GWAS. *BMC Genomics*, *21*(1), 1–13.
- Cheng, Z., Hou, D., Ge, W., Li, X., Xie, L., Zheng, H., et al. (2020). Integrated mRNA, MicroRNA transcriptome and degradome analyses provide insights into stamen development in Moso Bamboo. *Plant & Cell Physiology*, *61*(1), 76–87.
- Colaneri, A. C., & Jones, A. M. (2013). Genome-wide quantitative identification of DNA differentially methylated sites in Arabidopsis seedlings growing at different water potential. *PLoS One*, *8*(4), e59878.
- Collard, B. C., Jahufer, M., Brouwer, J., & Pang, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, *142*(1), 169–196.
- Coruh, C., Shahid, S., & Axtell, M. J. (2014). Seeing the forest for the trees: annotating small RNA producing genes in plants. *Current Opinion in Plant Biology*, *18*, 87–95.
- Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., et al. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, *16*(1), 123–131.
- Dai, X., Zhuang, Z., & Zhao, P. X. (2018). psRNATarget: A plant small RNA target analysis server (2017 release). *Nucleic Acids Research*, *46*(W1), W49–W54.
- Das, P., Nutan, K. K., Singla-Pareek, S. L., & Pareek, A. (2015). Understanding salinity responses and adopting 'omics-based' approaches to generate salinity tolerant cultivars of rice. *Frontiers in Plant Science*, *6*, 712.
- Deinlein, U., Stephan, A. B., Horie, T., Luo, W., Xu, G., & Schroeder, J. I. (2014). Plant salt-tolerance mechanisms. *Trends in Plant Science*, *19*(6), 371–379.

- van Dijk, K. V., Ding, Y., Malkaram, S. A., Riethoven, J.-J., Liu, R., Yang, J., et al. (2010). Dynamic changes in genome-wide histone H3 changes Lysine 4 methylation patterns in response to dehydration stress in *Arabidopsis thaliana*. *BMC Plant Biology*, *10*, 238.
- Do, P. T., Hoang, Y. H., Le, M. Q., Tang, H. T., & Nguyen, D. H. (2018). OsHKT1; 3 gene sequence polymorphisms and expression profile in rice (*Oryza sativa* L.). *African Journal of Agricultural Research*, *13*(46), 2659–2667.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21.
- Duan, P., Ni, S., Wang, J., Zhang, B., Xu, R., Wang, Y., et al. (2015). Regulation of OsGRF4 by OsmiR396 controls grain size and yield in rice. *Nature. Plants*, *2*(1), 1–5.
- Dubin, M. J., Zhang, P., Meng, D., Remigereau, M.-S., Osborne, E. J., Casale, F. P., et al. (2015). DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. *elife*, *4*, e05255.
- Dwivedi, S. L., Scheben, A., Edwards, D., Spillane, C., & Ortiz, R. (2017). Assessing and exploiting functional diversity in germplasm pools to enhance abiotic stress adaptation and yield in cereals and food legumes. *Frontiers in Plant Science*, *8*(1461).
- Ehrenreich, I. M., & Purugganan, M. D. (2008). Sequence variation of MicroRNAs and their binding sites in *Arabidopsis*. *Plant Physiology*, *146*(4), 1974–1982.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, *6*(5), e19379.
- Fahlgren, N., & Carrington, J. C. (2010). miRNA target prediction in plants. *Plant MicroRNAs: Springer*, 51–57.
- Fahlgren, N., Howell, M. D., Kasschau, K. D., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., et al. (2007). High-throughput sequencing of *Arabidopsis* microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLoS One*, *2*(2), e219.
- Fahlgren, N., Sullivan, C. M., Kasschau, K. D., Chapman, E. J., Cumbie, J. S., Montgomery, T. A., et al. (2009). Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA (New York, N.Y.)*, *15*(5), 992–1002.
- Famoso, A. N., Zhao, K., Clark, R. T., Tung, C. W., Wright, M. H., Bustamante, C., et al. (2011). Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. *PLoS Genetics*, *7*(8), e1002221.
- Fan, K.-H., Devos, K. M., & Schliekelman, P. (2020). Strategies for eQTL mapping in allopolyploid organisms. *Theoretical and Applied Genetics*, *133*, 2477–2497.
- Feltus, F. A., Wan, J., Schulze, S. R., Estill, J. C., Jiang, N., & Paterson, A. H. (2004). An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Research*, *14*(9), 1812–1819.
- Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, *7*(6), 461.
- Fu, L., Shen, Q., Kuang, L., Wu, D., & Zhang, G. (2019). Transcriptomic and alternative splicing analyses reveal mechanisms of the difference in salt tolerance between barley and rice. *Environmental and Experimental Botany*, *166*, 103810.
- Furini, A., Koncz, C., Salamini, F., & Bartels, D. (1997). High level transcription of a member of a repeated gene family confers dehydration tolerance to callus tissue of *Craterostigma plantagineum*. *The EMBO Journal*, *16*(12), 3599–3608.
- Gallego-Bartolomé, J., Liu, W., Kuo, P. H., Feng, S., Ghoshal, B., Gardiner, J., et al. (2019). Co-targeting RNA polymerases IV and V promotes efficient de novo DNA methylation in *Arabidopsis*. *Cell*, *176*(5), 1068–1082, e19.
- Galpaz, N., Gonda, I., Shem-Tov, D., Barad, O., Tzuri, G., Lev, S., et al. (2018). Deciphering genetic factors that determine melon fruit-quality traits using RNA-Seq-based high-resolution QTL and eQTL mapping. *The Plant Journal*, *94*(1), 169–191.
- Ganie, S. A. (2020). RNA chaperones: Potential candidates for engineering salt tolerance in rice. *Crop Science*, *60*(2), 530–540.
- Giresi, P. G., Kim, J., McDaniel, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885.
- Gopalakrishnan, T., Hasan, M. K., Haque, A. T. M. S., Jayasinghe, S. L., & Kumar, L. (2019). Sustainability of coastal agriculture under climate change. *Sustainability*, *11*(24), 7200.
- Goswami, K., Mittal, D., Gautam, B., Sopory, S. K., & Sanan-Mishra, N. (2020). Mapping the salt stress-induced changes in the root miRNome in Pokkali rice. *Biomolecules*, *10*(4), 498.
- Gray, S. B., & Brady, S. M. (2016). Plant developmental responses to climate change. *Developmental Biology*, *419*(1), 64–77.
- Gregorio, G., Islam, M., Vergara, G., & Thirumeni, S. (2013). Recent advances in rice science to design salinity and other abiotic stress tolerant rice varieties. *SABRAO Journal of Breeding and Genetics*, *45*(1), 31–40.
- Guo, T., Yang, J., Li, D., Sun, K., Luo, L., Xiao, W., et al. (2019). Integrating GWAS, QTL, mapping and RNA-seq to identify candidate genes for seed vigor in rice (*Oryza sativa* L.). *Molecular Breeding*, *39*(6), 87.
- Halfter, U., Ishitani, M., & Zhu, J.-K. (2000). The *Arabidopsis* SOS2 protein kinase physically interacts with and is activated by the calcium-binding protein SOS3. *Proceedings of the National Academy of Sciences*, *97*(7), 3735–3740.
- Hansen, B. G., Halkier, B. A., & Kliebenstein, D. J. (2008). Identifying the molecular basis of QTLs: eQTLs add a new dimension. *Trends in Plant Science*, *13*(2), 72–77.
- Haque T., Elias S.M., Razzaque S., Biswas S., Khan S.F., Jewel G.N.A., et al. (2020). Natural variation in growth and physiology under salt stress in rice: QTL mapping in a Horkuch × IR29 mapping population at seedling and reproductive stages. *bioRxiv*.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589.
- Henderson, I. R., & Jacobsen, S. E. (2007). Epigenetic inheritance in plants. *Nature*, *447*(7143), 418–424.

- Herbert, L., Meunier, A.-C., Bes, M., Vernet, A., Portefaix, M., Durand, F., et al. (2020). Beyond seek and destroy: How to generate allelic series using genome editing tools. *Rice (N Y)*, *13*(1), 5, -.
- Horie, T., Hauser, F., & Schroeder, J. I. (2009). HKT transporter-mediated salinity resistance mechanisms in Arabidopsis and monocot crop plants. *Trends in Plant Science*, *14*(12), 660–668.
- Huang, S., Xin, S., Xie, G., Han, J., Liu, Z., Wang, B., et al. (2020). Mutagenesis reveals that the rice OsMPT3 gene is an important osmotic regulatory factor. *The Crop Journal*, *8*(3), 465–479.
- Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, *42*(11), 961–967.
- Jaccoud, D., Peng, K., Feinstein, D., & Kilian, A. (2001). Diversity arrays: A solid state technology for sequence information independent genotyping. *Nucleic Acids Research*, *29*(4), e25, -e.
- Jadamba, C., Kang, K., Paek, N.-C., Lee, S. I., & Yoo, S.-C. (2020). Overexpression of rice expansin7 (Osexpa7) confers enhanced tolerance to salt stress in rice. *International Journal of Molecular Sciences*, *21*(2), 454.
- Jansen, R. C., & Nap, J.-P. (2001). Genetical genomics: The added value from segregation. *Trends in Genetics*, *17*(7), 388–391.
- Jiao, Y., Wang, Y., Xue, D., Wang, J., Yan, M., Liu, G., et al. (2010). Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nature Genetics*, *42*(6), 541.
- Jofuku, K. D., Den Boer, B., Van Montagu, M., & Okamoto, J. K. (1994). Control of Arabidopsis flower and seed development by the homeotic gene APETALA2. *The Plant Cell*, *6*(9), 1211–1225.
- Johnson, N. R., Yeoh, J. M., Coruh, C., & Axtell, M. J. (2016). Improved placement of multi-mapping small RNAs. *G3: Genes| Genomes| Genetics*, *6*(7), 2103–2111.
- Jones-Rhoades, M. W., Bartel, D. P., & Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology*, *57*, 19–53.
- Kakrana, A., Hammond, R., Patel, P., Nakano, M., & Meyers, B. C. (2014). sPARTA: A parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Research*, *42*(18), e139.
- Kazazian, H. H., Wong, C., Yousoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, *332*(6160), 164–166.
- Kent, N. A., Adams, S., Moorhouse, A., & Paszkiewicz, K. (2011). Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation DNA sequencing. *Nucleic Acids Research*, *39*(5), e26, -e.
- Kliebenstein, D. (2009). Quantitative genomics: Analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology*, *60*, 93–114.
- Kobayashi, N. I., Yamaji, N., Yamamoto, H., Okubo, K., Ueno, H., Costa, A., et al. (2017). OsHKT1; 5 mediates Na⁺ exclusion in the vasculature to protect leaf blades and reproductive tissues from salt toxicity in rice. *The Plant Journal*, *91*(4), 657–670.
- Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., et al. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genetics*, *5*(7), e1000551.
- Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, *42*(D1), D68–D73.
- Kuang, Z., Wang, Y., Li, L., & Yang, X. (2019). miRDeep-P2: Accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics (Oxford, England)*, *35*(14), 2521–2522.
- Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., Langford, K. W., Nova, I. C., Samson, J. M., et al. (2013). Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences*, *110*(47), 18904–18909.
- Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews. Genetics*, *11*(3), 204–220.
- Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, *128*(4), 707–719.
- Li, R., Jeong, K., Davis, J. T., Kim, S., Lee, S., Michelmore, R. W., et al. (2018). Integrated QTL and eQTL mapping provides insights and candidate genes for fatty acid composition, flowering time, and growth traits in a F2 population of a novel synthetic allopolyploid Brassica napus. *Frontiers in plant science*, *9*, 1632.
- Li, W., Pang, S., Lu, Z., & Jin, B. (2020). Function and mechanism of WRKY transcription factors in abiotic stress responses of plants. *Plants (Basel)*, *9*(11).
- Li, Z., Wang, P., You, C., Yu, J., Zhang, X., Yan, F., et al. (2020). Combined GWAS and eQTL analysis uncovers a genetic regulatory network orchestrating the initiation of secondary cell wall development in cotton. *New Phytologist*, *226*(6), 1738–1752.
- Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930.
- Lin, S., Sasaki, T., & Yano, M. (1998). Mapping quantitative trait loci controlling seed dormancy and heading date in rice, *Oryza sativa* L., using backcross inbred lines. *Theoretical and Applied Genetics*, *96*(8), 997–1003.
- Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews. Genetics*, *14*(1), 49–61.
- Liu, B. H. (2017). *Statistical genomics: linkage, mapping, and QTL analysis*. CRC press.
- Liu, J., Ishitani, M., Halfter, U., Kim, C.-S., & Zhu, J.-K. (2000). The Arabidopsis thaliana SOS2 gene encodes a protein kinase that is required for salt tolerance. *Proceedings of the national academy of sciences*, *97*(7), 3730–3734.
- Liu, Q., Wang, H., Zhu, L., Hu, H., & Sun, Y. (2013). Genome-wide identification and analysis of miRNA-related single nucleotide polymorphisms (SNPs) in rice. *Rice*, *6*(1), 1–10.

- Liu, Q., Yang, T., Yu, T., Zhang, S., Mao, X., Zhao, J., et al. (2017). Integrating small RNA sequencing with QTL mapping for identification of miRNAs and their target genes associated with heat tolerance at the flowering stage in rice. *Frontiers in Plant Science*, 8, 43.
- Liu, T., Fang, C., Ma, Y., Shen, Y., Li, C., Li, Q., et al. (2016). Global investigation of the co-evolution of MIRNA genes and micro RNA targets during soybean domestication. *The Plant Journal*, 85(3), 396–409.
- Liu, Y., Teng, C., Xia, R., & Meyers, B. C. (2020). PhasiRNAs in plants: Their biogenesis, genic sources, and roles in stress responses, development, and reproduction. *The Plant Cell*, 32(10), 3059–3080.
- Llave, C., Kasschau, K. D., Rector, M. A., & Carrington, J. C. (2002). Endogenous and silencing-associated small RNAs in plants. *The Plant Cell*, 14(7), 1605–1619.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
- Lovell, J. T., Jenkins, J., Lowry, D. B., Mamidi, S., Sreedasyam, A., Weng, X., et al. (2018). The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nature Communications*, 9(1), 5213.
- Lu, Z., Hofmeister, B. T., Vollmers, C., DuBois, R. M., & Schmitz, R. J. (2017). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Research*, 45(6), e41, -e.
- Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., & Schmitz, R. J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature plants*, 5(12), 1250–1259.
- Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648), 251–260.
- Ma, Z., Coruh, C., & Axtell, M. J. (2010). Arabidopsis lyrata small RNAs: transient MIRNA and small interfering RNA loci within the Arabidopsis genus. *The Plant Cell Online*, 22(4), 1090–1103.
- Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., et al. (2018). Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *The Plant Cell*, 30(1), 15–36.
- Margarido, G. R., Souza, A. P., & Garcia, A. A. (2007). OneMap: Software for genetic mapping in outcrossing species. *Hereditas*, 144(3), 78–79.
- Martin, R. C., Liu, P.-P., Goloviznina, N. A., & Nonogaki, H. (2010). microRNA, seeds, and Darwin?: Diverse function of miRNA in seed biology and plant responses to stress. *Journal of Experimental Botany*, 61(9), 2229–2234.
- Matzke, M. A., Kanno, T., & Matzke, A. J. M. (2015). RNA-directed DNA methylation: The evolution of a complex epigenetic pathway in flowering plants. *Annual Review of Plant Biology*, 66, 243–267.
- McCouch, S. R., Teytelman, L., Xu, Y., Lobos, K. B., Clare, K., Walton, M., et al. (2002). Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Research*, 9(6), 199–207.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303.
- de Meaux, J., Hu, J.-Y., Tartler, U., & Goebel, U. (2008). Structurally different alleles of the ath-MIR824 microRNA precursor are maintained at high frequency in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 105(26), 8994–8999.
- Meyer, E., Aglyamova, G., & Matz, M. (2011). Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Molecular Ecology*, 20(17), 3599–3616.
- Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., et al. (2008). Criteria for annotation of plant MicroRNAs. *The Plant Cell Online*, 20(12), 3186–3190.
- Mickelbart, M. V., Hasegawa, P. M., & Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nature Reviews. Genetics*, 16(4), 237–251.
- Miura, K., Ikeda, M., Matsubara, A., Song, X.-J., Ito, M., Asano, K., et al. (2010). OsSPL14 promotes panicle branching and higher grain productivity in rice. *Nature Genetics*, 42(6), 545–549.
- Mohorianu, I., Stocks, M. B., Applegate, C. S., Folkes, L., & Moulton, V. (2017). *The UEA small RNA workbench: A suite of computational tools for small RNA analysis. MicroRNA detection and target identification* (pp. 193–224). Springer.
- Mondini, L., Nachit, M., Porceddu, E., & Pagnotta, M. A. (2012). Identification of SNP mutations in DREB1, HKT1, and WRKY1 genes involved in drought and salt stress tolerance in durum wheat (*Triticum turgidum* L. var durum). *OmicS: A Journal of Integrative Biology*, 16(4), 178–187.
- Montes, R. A. C., De Paoli, E., Accerbi, M., Rymarquis, L. A., Mahalingam, G., Marsch-Martínez, N., et al. (2014). Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nature Communications*, 5.
- Morgado, L., & Johannes, F. (2019). Computational tools for plant small RNA detection and categorization. *Briefings in Bioinformatics*, 20(4), 1181–1192.
- Mäki-Tanila, A., & Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1), 355–367.
- Nagarajan, Y., Rongala, J., Luang, S., Singh, A., Shadiac, N., Hayes, J., et al. (2016). A barley efflux transporter operates in a Na⁺-dependent manner, as revealed by a multidisciplinary platform. *The Plant Cell*, 28(1), 202–218.
- Nair, S. K., Wang, N., Turuspekov, Y., Pourkheirandish, M., Sinsuwongwat, S., Chen, G., et al. (2010). Cleistogamous flowering in barley arises from the suppression of microRNA-guided HvAP2 mRNA cleavage. *Proceedings of the National Academy of Sciences*, 107(1), 490–495.
- Ni, P., Huang, N., Nie, F., Zhang, J., Zhang, Z., Wu, B., et al. (2021). Genome-wide detection of cytosine methylations in plant from nanopore sequencing data using deep learning. *Nature Communications*, 12, 2021.02.07.430077.
- Oi, T., Enomoto, S., Nakao, T., Arai, S., Yamane, K., & Taniguchi, M. (2019). Three-dimensional ultrastructural change of chloroplasts in rice mesophyll cells responding to salt stress. *Annals of Botany*, 125(5), 833–840.

- Osman, K. A., Tang, B., Wang, Y., Chen, J., Yu, F., Li, L., et al. (2013). Dynamic QTL analysis and candidate gene mapping for waterlogging tolerance at maize seedling stage. *PLoS One*, *8*(11), e79305.
- Pandey, P., Irulappan, V., Bagavathiannan, M. V., & Senthil-Kumar, M. (2017). Impact of combined abiotic and biotic stresses on plant growth and avenues for crop improvement by exploiting physio-morphological traits. *Frontiers in Plant Science*, *8*, 537.
- Park, S.-Y., & Kim, J.-S. (2020). A short guide to histone deacetylases including recent progress on class II enzymes. *Experimental & Molecular Medicine*, *52*(2), 204–212.
- Parvathaneni R.K., Kumar I., Braud M., Eveland A.L. (2020). Regulatory signatures of drought response in stress resilient *Sorghum bicolor*. *bioRxiv*.
- Pazos-Navarro, M., Castello, M., Bennett, R. G., Nichols, P., & Croser, J. (2017). In vitro-assisted single-seed descent for breeding-cycle compression in subterranean clover (*Trifolium subterraneum* L.). *Crop and Pasture Science (New York, N.Y.)*, *68*(11), 958–966.
- Pecinka, A., Dinh, H. Q., Baubec, T., Rosa, M., Lettner, N., & Scheid, O. M. (2010). Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in *Arabidopsis*. *The Plant Cell*, *22*(9), 3118–3129.
- Peng, T., Teotia, S., Tang, G., & Zhao, Q. (2019). MicroRNAs meet with quantitative trait loci: Small powerful players in regulating quantitative yield traits in rice. *Wiley Interdiscip Rev RNA*, *10*(6), e1556.
- Perkel, J. M. (2021). Five reasons why researchers should learn to love the command line. *Nature*, *590*(7844), 173–174.
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*(5), e37135.
- Plongthongkum, N., Diep, D. H., & Zhang, K. (2014). Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature Reviews. Genetics*, *15*(10), 647–661.
- Princi, M. P., Lupini, A., Araniti, F., Longo, C., Mauceri, A., Sunseri, F., et al. (2016). Boron toxicity and tolerance in plants: Recent advances and future perspectives. *Plant Metal Interaction*, 115–147.
- Rajkumar, M. S., Gupta, K., Khemka, N. K., Garg, R., & Jain, M. (2020). DNA methylation reprogramming during seed development and its functional relevance in seed size/weight determination in chickpea. *Communications Biology*, *3*(1), 1–13.
- Razzaque, S., Elias, S. M., Haque, T., Biswas, S., Jewel, G. M. N. A., Rahman, S., et al. (2019). Gene Expression analysis associated with salt stress in a reciprocally crossed rice population. *Scientific Reports*, *9*(1), 8249.
- Razzaque, S., Haque, T., Elias, S. M., Rahman, M. S., Biswas, S., Schwartz, S., et al. (2017). Reproductive stage physiological and transcriptional responses to salinity stress in reciprocal populations derived from tolerant (Horkuch) and susceptible (IR29) rice. *Scientific Reports*, *7*(1), 46138.
- Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., & Bartel, D. P. (2002). MicroRNAs in plants. *Genes & Development*, *16*(13), 1616–1626.
- Ren, Z.-H., Gao, J.-P., Li, L.-G., Cai, X.-L., Huang, W., Chao, D.-Y., et al. (2005). A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nature Genetics*, *37*(10), 1141–1146.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140.
- Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouzé, P., & Van de Peer, Y. (2003). Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiology*, *132*(3), 1162–1176.
- Rushton, P. J., Somssich, I. E., Ringler, P., & Shen, Q. J. (2010). WRKY transcription factors. *Trends in Plant Science*, *15*(5), 247–258.
- Sasaki, T. (2005). The map-based sequence of the rice genome. *Nature*, *436*(7052), 793–800.
- Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*(2), 149–161.
- Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews. Genetics*, *19*(6), 329–346.
- Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)*, *28*(10), 1353–1358.
- Shahid, S., & Axtell, M. J. (2013). Identification and annotation of small RNA genes using ShortStack. *Methods (San Diego, Calif.)*.
- Shen, Y.-J., Jiang, H., Jin, J.-P., Zhang, Z.-B., Xi, B., He, Y.-Y., et al. (2004). Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiology*, *135*(3), 1198–1205.
- Shi, W., Liu, D., Hao, L., Wu, C.-a, Guo, X., & Li, H. (2014). GhWRKY39, a member of the WRKY transcription factor family in cotton, has a positive role in disease resistance and salt stress tolerance. *Plant Cell, Tissue and Organ Culture (PCTOC)*, *118*(1), 17–32.
- Shohan, M. U. S., Sinha, S., Nabila, F. H., Dastidar, S. G., & Seraj, Z. I. (2019). HKT1; 5 transporter gene expression and association of amino acid substitutions with salt tolerance across rice genotypes. *Frontiers in Plant Science*, *10*, 1420.
- Simon, S. A., Zhai, J., Nandety, R. S., McCormick, K. P., Zeng, J., Mejia, D., et al. (2009). Short-read sequencing technologies for transcriptional analyses. *Annual Review of Plant Biology*, *60*, 305–333.
- Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, *14*(4), 407.
- Somasundaram, S., Véry, A.-A., Vinekar, R. S., Ishikawa, T., Kumari, K., Pulipati, S., et al. (2020). Homology modeling identifies crucial amino-acid residues that confer higher Na⁺ transport capacity of OeHKT1; 5 from *Oryza coarctata* Roxb. *Plant and Cell Physiology*, *61*(7), 1321–1334.
- Song, X., Li, Y., Cao, X., & Qi, Y. (2019). MicroRNAs and their regulatory roles in plant–environment interactions. *Annual Review of Plant Biology*, *70*, 489–525.
- Soppe, W. J., Jacobsen, S. E., Alonso-Blanco, C., Jackson, J. P., Kakutani, T., Koornneef, M., et al. (2000). The late flowering phenotype of FWA mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Molecular Cell*, *6*(4), 791–802.

- Srisikantharajah, K., Osumi, S., Chuamnakhong, S., Nampei, M., Amas, J. C., Gregorio, G. B., et al. (2020). Contribution of two different Na⁺ transport systems to acquired salinity tolerance in rice. *Plant Science*, 297, 110517.
- Swetha, C., Basu, D., Pachamuthu, K., Tirumalai, V., Nair, A., Prasad, M., et al. (2018). Major domestication-related phenotypes in indica rice are due to loss of miRNA-mediated laccase silencing. *The Plant Cell*, 30(11), 2649–2662.
- Tanksley, S. D. (1993). Mapping polygenes. *Annual Review of Genetics*, 27(1), 205–233.
- Tao, X., Feng, S., Zhao, T., & Guan, X. (2020). Efficient chromatin profiling of H3K4me3 modification in cotton using CUT&Tag. *Plant Methods*, 16(1), 1–15.
- Tarbell, E. D., & Liu, T. (2019). HMMRATAC: A Hidden Markov Modeler for ATAC-seq. *Nucleic Acids Research*, 47(16), e91, -e.
- Taylor, R. S., Tarver, J. E., Foroozani, A., & Donoghue, P. C. J. (2017). MicroRNA annotation of plant genomes- Do it right or not at all. *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology*, 39(2), 1600113.
- Temnykh, S., Park, W. D., Ayres, N., Cartinhour, S., Hauck, N., Lipovich, L., et al. (2000). Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theoretical and Applied Genetics*, 100(5), 697–712.
- Thatcher, S. R., Burd, S., Wright, C., Lers, A., & Green, P. J. (2015). Differential expression of miRNAs and their target genes in senescing leaves and siliques: Insights from deep sequencing of small RNAs and cleaved target RNAs. *Plant, Cell & Environment*, 38(1), 188–200.
- Thody, J., Folkes, L., Medina-Calzada, Z., Xu, P., Dalmay, T., & Moulton, V. (2018). PAREsnip2: A tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules. *Nucleic Acids Research*, 46(17), 8730–8739.
- Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding and Biotechnology*, 2(3), 195–212.
- Todesco, M., Balasubramanian, S., Cao, J., Ott, F., Sureshkumar, S., Schneeberger, K., et al. (2012). Natural variation in biogenesis efficiency of individual *Arabidopsis thaliana* microRNAs. *Current Biology*, 22(2), 166–170.
- Tomás, D., Brazao, J., Viegas, W., & Silva, M. (2013). Differential effects of high-temperature stress on nuclear topology and transcription of repetitive noncoding and coding rye sequences. *Cytogenetic and Genome Research*, 139(2), 119–127.
- Torti, S., Schlesier, R., Thümler, A., Bartels, D., Römer, P., Koch, B., et al. (2021). Transient reprogramming of crop plants for agronomic performance. *Nature Plants*, 7(2), 159–171.
- Tricker, P. J., Gibbins, J. G., Rodríguez López, C. M., Hadley, P., & Wilkinson, M. J. (2012). Low relative humidity triggers RNA-directed de novo DNA methylation and suppression of genes controlling stomatal development. *Journal of Experimental Botany*, 63(10), 3799–3813.
- Tse, O. Y. O., Jiang, P., Cheng, S. H., Peng, W., Shang, H., Wong, J., et al. (2021). Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 118(5).
- Tseng, K.-C., Chiang-Hsieh, Y.-F., Pai, H., Chow, C.-N., Lee, S.-C., Zheng, H.-Q., et al. (2018). microRPM: A microRNA prediction model based only on plant small RNA sequencing data. *Bioinformatics (Oxford, England)*, 34(7), 1108–1115.
- Turco, G., Schnable, J. C., Pedersen, B., & Freeling, M. (2013). Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. *Frontiers in Plant Science*, 4, 170.
- Van de Velde, J., Van Bel, M., Vanechoutte, D., & Vandepoele, K. (2016). A collection of conserved noncoding sequences to study gene regulation in flowering plants. *Plant Physiology*, 171(4), 2586–2598.
- Wang, C., Ye, J., Tang, W., Liu, Z., Zhu, C., Wang, M., et al. (2013). Loop nucleotide polymorphism in a putative miRNA precursor associated with seed length in rice (*Oryza sativa* L.). *International Journal of Biological Sciences*, 9(6), 578.
- Wang, J., Nan, N., Li, N., Liu, Y., Wang, T.-J., Hwang, I., et al. (2020). A DNA Methylation Reader–Chaperone Regulator–Transcription Factor Complex Activates OsHKT1; 5 expression during salinity stress. *The Plant Cell*, 32(11), 3535–3558.
- Wang, J., Yu, H., Xie, W., Xing, Y., Yu, S., Xu, C., et al. (2010). A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. *The Plant Journal*, 63(6), 1063–1074.
- Wang, J., Zhu, J., Zhang, Y., Fan, F., Li, W., Wang, F., et al. (2018). Comparative transcriptome analysis reveals molecular response to salinity stress of salt-tolerant and sensitive genotypes of indica rice at seedling stage. *Scientific Reports*, 8(1), 1–13.
- Wang, M., Qin, L., Xie, C., Li, W., Yuan, J., Kong, L., et al. (2014). Induced and constitutive DNA methylation in a salinity-tolerant wheat introgression line. *Plant and Cell Physiology*, 55(7), 1354–1365.
- Wang, W.-S., Pan, Y.-J., Zhao, X.-Q., Dwivedi, D., Zhu, L.-H., Ali, J., et al. (2011). Drought-induced site-specific DNA methylation and its association with drought tolerance in rice (*Oryza sativa* L.). *Journal of Experimental Botany*, 62(6), 1951–1960.
- Wang, X., Weigel, D., & Smith, L. M. (2013). Transposon variants and their effects on gene expression in *Arabidopsis*. *PLoS Genetics*, 9(2), e1003255.
- Wang, Y., Shen, D., Bo, S., Chen, H., Zheng, J., Zhu, Q.-H., et al. (2010). Sequence variation and selection of small RNAs in domesticated rice. *BMC Evolutionary Biology*, 10(1), 1–10.
- Wang, Z., & Baulcombe, D. C. (2020). Transposon age and non-CG methylation. *Nature Communications*, 11(1), 1–9.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1), 57–63.
- Wen, M., Xie, M., He, L., Wang, Y., Shi, S., & Tang, T. (2016). Expression variations of miRNAs and mRNAs in rice (*Oryza sativa*). *Genome Biology and Evolution*, 8(11), 3529–3544.
- Wendte, J. M., & Pikaard, C. S. (2017). The RNAs of RNA-directed DNA methylation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1860(1), 140–148.
- Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., et al. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *The Plant Cell*, 28(10), 2365–2384.

- Xu, B., Waters, S., Byrt, C. S., Plett, D., Tyerman, S. D., Tester, M., et al. (2018). Structural variations in wheat HKT1;5 underpin differences in Na (+) transport capacity. *Cellular and Molecular Life Sciences: CMLS*, 75(6), 1133–1144.
- Yang, M., Woolfenden, H. C., Zhang, Y., Fang, X., Liu, Q., Vigh, M. L., et al. (2020). Intact RNA structurome reveals mRNA structure-mediated regulation of miRNA cleavage in vivo. *Nucleic Acids Research*, 48(15), 8767–8781.
- Yao, Z., You, F. M., N'Diaye, A., Knox, R. E., McCartney, C., Hiebert, C. W., et al. (2020). Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics*, 21(1), 1–16.
- Yeo, B. P. H., Bhave, M., & San Hwang, S. (2018). Effects of acute salt stress on modulation of gene expression in a Malaysian salt-tolerant indigenous rice variety, Bajong. *Journal of Plant Research*, 131(1), 191–202.
- You, C., Cui, J., Wang, H., Qi, X., Kuo, L.-Y., Ma, H., et al. (2017). Conservation and divergence of small RNA pathways and microRNAs in land plants. *Genome Biology*, 18(1), 1–19.
- Zemach, A., Kim, M. Y., Silva, P., Rodrigues, J. A., Dotson, B., Brooks, M. D., et al. (2010). Local DNA hypomethylation activates genes in rice endosperm. *Proceedings of the National Academy of Sciences*, 107(43), 18729–18734.
- Zhang, B. (2015). MicroRNA: A new target for improving plant tolerance to abiotic stress. *Journal of Experimental Botany*, 66(7), 1749–1761.
- Zhang, B. H., Pan, X. P., Wang, Q. L., George, P. C., & Anderson, T. A. (2005). Identification and characterization of new plant microRNAs using EST analysis. *Cell Research*, 15(5), 336–360.
- Zhang, H., Lang, Z., & Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. *Nature Reviews. Molecular Cell Biology*, 19(8), 489–506.
- Zhang, L., Peng, Y., Wei, X., Dai, Y., Yuan, D., Lu, Y., et al. (2014). Small RNAs as important regulators for the hybrid vigour of super-hybrid rice. *Journal of Experimental Botany*, 65(20), 5989–6002.
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoutte, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9), 1–9.
- Zhang, Y., Malzahn, A. A., Sretenovic, S., & Qi, Y. (2019). The emerging and uncultivated potential of CRISPR technology in plant science. *Nature Plants*, 5(8), 778–794.
- Zhang, Y.-M., & Xu, S. (2004). Mapping quantitative trait loci in F2 incorporating phenotypes of F3 progeny. *Genetics*, 166(4), 1981–1993.
- Zhao, C., Zhang, H., Song, C., Zhu, J.-K., & Shabala, S. (2020). Mechanisms of plant responses and adaptation to soil salinity. *The Innovation*, 1(1), 100017.
- Zhao, K., Wright, M., Kimball, J., Eizenga, G., McClung, A., Kovach, M., et al. (2010). Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS One*, 5(5), e10780.
- Zhao, M., Liu, B., Wu, K., Ye, Y., Huang, S., Wang, S., et al. (2015). Regulation of OsmiR156h through alternative polyadenylation improves grain yield in rice. *PLoS One*, 10(5), e0126154.
- Zheng, C., & Hayes, J. J. (2003). Structures and interactions of the core histone tail domains. *Biopolymers: Original Research on Biomolecules*, 68(4), 539–546.
- Zhong, Z., Feng, S., Duttke, S. H., Potok, M. E., Zhang, Y., Gallego-Bartolomé, J., et al. (2021). DNA methylation-linked chromatin accessibility affects genomic architecture in Arabidopsis. *Proceedings of the National Academy of Sciences*, 118(5).
- Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and prospects of association mapping in plants. *The Plant Genome*, 1(1), 5–20.
- Zilberman, D. (2017). An evolutionary case for functional gene body methylation in plants and animals. *Genome Biology*, 18(1), 1–3.

Section IV

Artificial intelligence and agribots

This page intentionally left blank

Deep Learning applied to computational biology and agricultural sciences

Renato Hidaka Torres¹, Fabricio Almeida Araujo¹, Edian Franklin Franco De Los Santos^{1,2}, Debmalya Barh³, Rommel Thiago Jucá Ramos¹ and Marcus de Barros Braga⁴

¹Federal University of Pará, Belém, Brazil, ²Santo Domingo Technological Institute, Santo Domingo, Dominican Republic, ³Institute of Integrative Omics and Applied Biotechnology, Purba Medinipur, West Bengal, India, ⁴Federal Rural University of Amazonia, Paragominas, Brazil

34.1 Introduction

The decrease in the cost of producing genomic data increased the importance of research in areas such as DNA sequencing, RNA sequencing, and high-throughput screening. This ended up creating new challenges in finding efficient ways to analyze data and provide insights into the function of biological systems. Computational biology uses knowledge of computer science, biology, statistics, chemistry, and engineering to analyze and infer the relationships between biological data and thus create computational solutions to address these issues. Bioinformatics research provides solutions on evolutionary aspects of molecular biology, metabolic pathways and networks, expression and regulation of genes and proteins, genomic annotation, and biomolecular interaction. Machine Learning methods are general-purpose approaches to learning functional data relationships without the need to define them a priori. Within computational biology, its application involves the ability to generate predictive models without the need for strong assumptions about mechanisms that are often unknown or insufficiently defined. Deep Neural Network (DNN), subarea of Machine Learning which assembles Deep Learning algorithms, takes the raw data in its lowest layer (input) and transforms them into representations of increasingly abstract features, successively combining the outputs of the previous layer in a data-oriented manner, encapsulating complex mathematical functions in the process. Deep Learning is now one of the most active fields in Machine Learning and shows to improve performance in image and speech recognition, understanding of natural language, and, more recently, in computational biology. The potential for using Deep Learning in high-performance biology is remarkable. It allows to better explore the availability of larger and larger datasets (e.g., DNA sequencing, RNA measurements, flow cytometry, or microscopy) by training complex networks with multiple layers that capture their internal. These deep networks, after being trained, are able to discover high-level features, improve the performance of traditional models, increase interpretability, and provide the understanding of the structure of biological data (Angermueller, Pärnamaa, Parts, & Stegle, 2016).

Likewise, Deep Learning has been used widely to solve historical challenges in agricultural sciences, such as in image processing and data analysis, with promising results and great potential. As Deep Learning has been successfully applied in a number of fields, more recently the turn to agriculture has come.

34.2 Deep Learning and Convolutional Neural Network

Machine Learning is an area of artificial intelligence, the objective of which implies the development of computational algorithms that are capable of transforming experience into expertise. In other words, Machine Learning algorithms aim to map patterns from an input domain to an output domain and subsequently recognize patterns from the output domain, even those not exemplified.

Fig. 34.1 illustrates the mentioned domain mapping process. In this figure the symbol f represents the Machine Learning model that is being built from examples of the input and output domains. After building the model, given a

sample of the input domain, it is expected that f be able to perform the proper mapping for the output domain. If this operation is satisfied, the Machine Learning model is said to have the ability to generalize.

A Machine Learning algorithm can be classified taking into account the adopted domain mapping methodology. Among the possible classifications, these algorithms can be classified as supervised and unsupervised algorithms. So, let X be a set of independent variables and Y a set of dependent variables, a dataset labeled L is defined by N pairs of input values and output values $(x_1, y_1), \dots, (x_N, y_N)$ where $x_i \in X$ and $y_i \in Y$. Thus the task of a supervised model implies learning the function $f: X \rightarrow Y$ from the set $L = (X_N, Y_N)$. After building the model, it is expected that the function $f: X \rightarrow Y$ be able to map values $(x \in X, y \in Y) \notin L$.

The development of supervised Machine Learning models implies the need for data labeling. However, in many real problems, labeling a sample of the input data is a complex or even unfeasible task. When this is the case, it is recommended to use unsupervised Machine Learning algorithms. In unsupervised Machine Learning the algorithm is built from unlabeled examples. In this type of algorithm the similarity between the input data is sought. That is, if the input data is similar, it is expected that this data will reflect in mapping for the same group as the output domain. In unsupervised problems the outgoing domain is built from clustering. Referring to Fig. 34.1, the training data of the unsupervised algorithm imply only the examples of the input domain. Because it is not labeled, the examples of the output domain are not used in training the model.

Among the Machine Learning algorithms in the literature, algorithms based on convolution are standing out in several areas. CNN (Convolutional Neural Network) is a Deep Learning approach that has been demonstrated efficiency in problems such as object classification and face, speak, and action recognition. Architectures such as LeNet-5, AlexNet, ZFNet, VGGNet, GoogleNet e ResNet show the evolution of CNNs and a tendency for increasingly deep architectures. According to Gu et al. (2018), CNNs with deep architecture can provide a better performance, since they increase the nonlinearity of the network. However, the depth also increases the complexity of the network and makes it difficult to optimize. So, regardless of architecture and depth, it is possible to notice that there is a pattern of the components used in a CNN. According to Goodfellow, Bengio, and Courville (2016), a convolutional layer is composed of the 3-upla <convolution, activation function, pooling>.

Convolution is a mathematical theorem applied to two functions f and g to obtain a third function h , defined as follows:

$$(f * g)(c) = h(c) = \sum_a f(a) \times g(c - a) \tag{34.1}$$

To illustrate, consider the scenario in Fig. 34.2. In this case a ball is thrown from a height h_1 and travels a distance A with probability $f(A)$. From the stop point the ball is relaunched from a height h_2 and travels a distance B , with

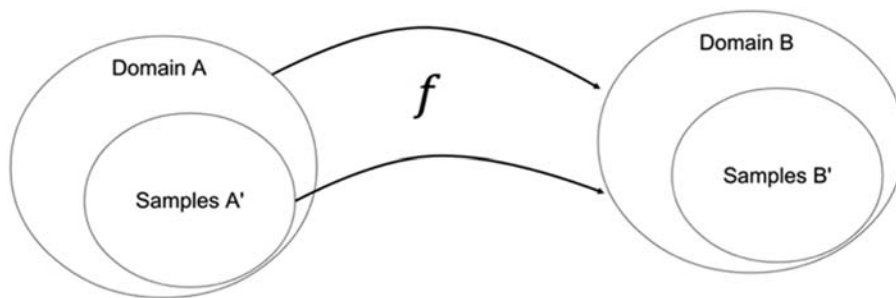


FIGURE 34.1 Domain mapping process performed by Machine Learning models.

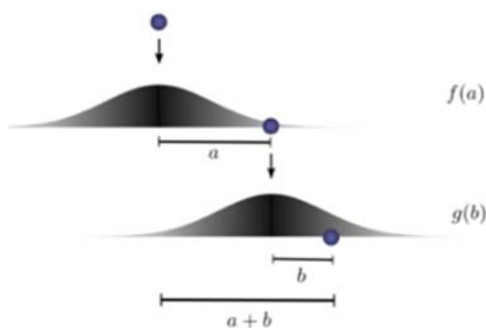


FIGURE 34.2 Convolution exemplified by the throw of the ball.

probability $g(B)$. Thus the total distance traveled by the ball is $c = a + b$. The probability of the ball traveling distance c is given by $f(A) \times g(B)$, that is, the probability of traveling distance A and the probability of traveling distance B .

Assuming $c = 8$, it is possible to obtain different values for a and b that satisfy the equation $a + b = 8$. Therefore to obtain the probability function of c , it is necessary to take into account all the possibilities of $f(A) \times g(B)$. Therefore the probability of the ball traveling a distance c can be defined as follows:

$$(f * g)(c) = h(c) = \sum_{a+b=c} f(a) \times g(b) \tag{34.2}$$

Assuming $b = c - a$, we have the convolution:

$$(f * g)(c) = h(c) = \sum_a f(a) \times g(c - a) \tag{34.3}$$

In the context of Artificial Neural Networks (ANNs), many applications use multidimensional input values. Applications that work with image manipulation are examples of networks that manipulate two-dimensional inputs. In this context the convolution is given by:

$$\begin{aligned} (f * g)(c_1, c_2) &= h(c_1, c_2) \\ &= \sum_{a_1, a_2} f(a_1, a_2) \times g(c_1 - a_1, c_2 - a_2) \end{aligned} \tag{34.4}$$

Two-dimensional convolution can be seen as a sliding window from one function to the other. In the case of the CNN, convolution implies the kernel function by sliding over the array of input values. Fig. 34.3 shows an example of a two-dimensional convolution performed by a CNN.

According to Buduma and Locascio (2017), the kernel functions of CNNs work as feature detectors. Each kernel has the function of learning specific characteristics of the input data. The learning of each kernel is given by adjusting the weights, during network training. In image classification or feature recognition problems, convolutional layers typically learn to detect specific edges, texture, shapes, and characteristics of the problem.

As an example, Fig. 34.4 presents an abstraction of the functioning of the convolutional layers. Each frame in Fig. 34.4 represents a kernel function performing a convolution.

According to Chollet (2017), convolutional kernels enable hierarchical learning. In this case the first convolution layer learns small local patterns, such as edges; the second layer of convolution learns more specific patterns, and so on. In the example in Fig. 34.4, the filters of the first convolutional layer learned generic forms of the object. In the second layer the combination of characteristics made it possible to learn in more specific ways.

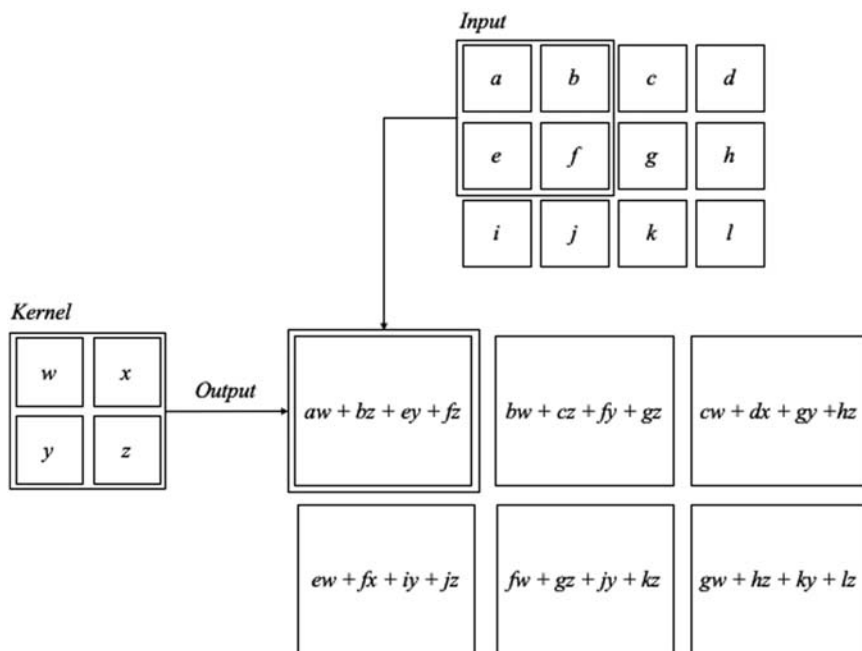


FIGURE 34.3 Example of two-dimensional convolution (Goodfellow et al., 2016).

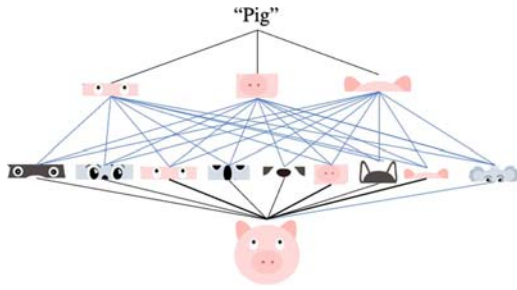


FIGURE 34.4 Abstraction of the CNN kernel functions (Chollet, 2017). CNN, Convolutional Neural Network.

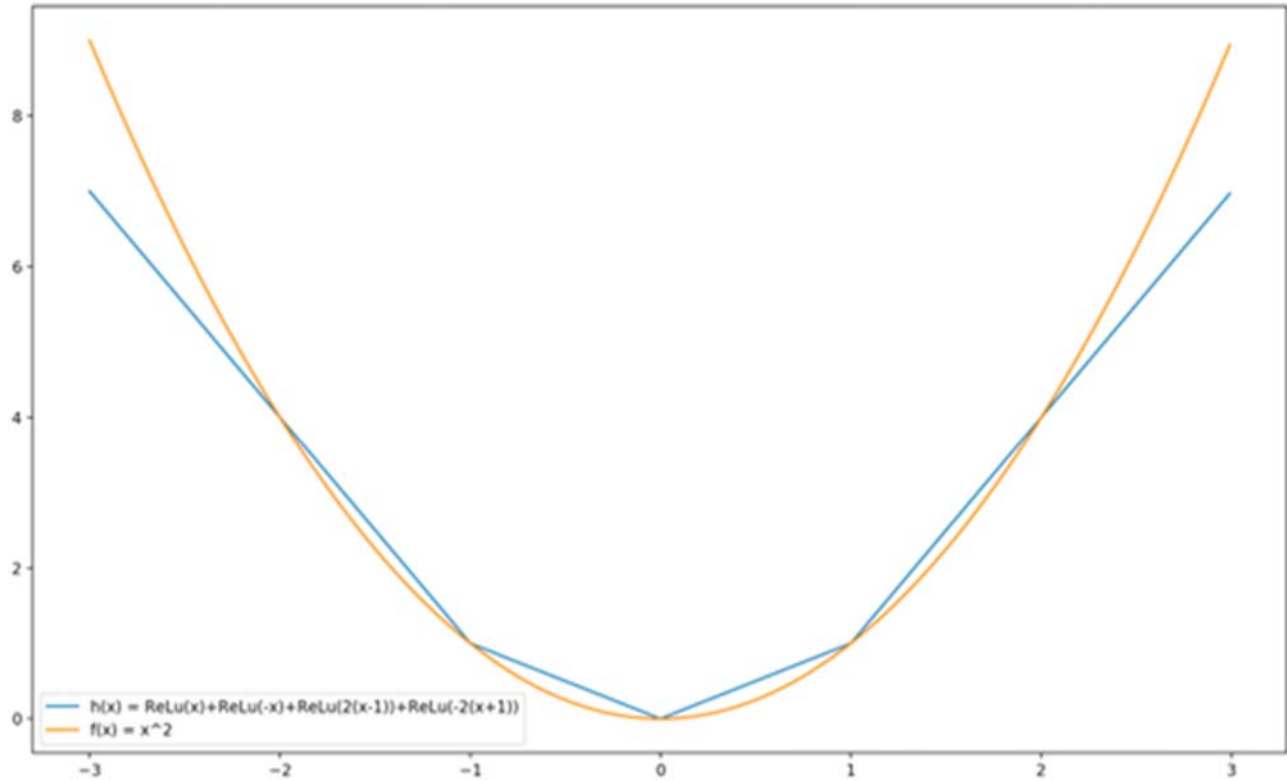


FIGURE 34.5 Approximation of the quadratic function by the ReLU function. *ReLU*, Rectified Linear Unit.

The convolution of a CNN is a linear system. In this sense, multiple convolutions also form a linear system. In order, for a CNN to solve nonlinearly separable problems, it is necessary to use nonlinear activation functions among each convolution layer. Studies such as Gu et al. (2018), Alcantara (2017), and LeCun, Bengio, and Hinton (2015) point out that the Rectified Linear Unit (ReLU) function and its variations have good performance for the neurons of the convolution layers. The ReLU function (see Eq. 34.2) calculates the maximum value of the input values and transforms all the negative input values to zero.

$$f(x) = \max(0, x) \quad (34.5)$$

According to Gu et al. (2018), the convergence of negative values allows the nonactivation of all neurons and, consequently, a sparse representation of the model. The sparse representation provided by the ReLU function allows CNNs to be trained efficiently, without the need for pretraining. In addition, the use of ReLU functions makes it possible to approach more complex functions. For example, consider the quadratic function $f(x) = x^2$. In this case we could use the following function ReLU, to approximate $f(x)$ (Fig. 34.5).

$$h(x) = \text{ReLU}(x) + \text{ReLU}(-x) + \text{ReLU}(2x - 2) + \text{ReLU}(-2x + 2) \quad (34.6)$$

The last component of 3-tuple is the pooling function. These functions analyze a set of neighbors and extract a characteristic that represents them. The goal is to merge semantically similar features and make feature mapping invariant, even if there are variations in input values. The Max Pooling function proposed by Zhou and Chellappa (1988) performs this representativity by extracting the maximum value among the observed neighbors. In addition to Max Pooling, Gu et al. (2018) emphasize that the functions LP Pooling, Mixed Pooling, Stochastic Pooling, Spectral Pooling, Spatial Pyramid Pooling, and Multiscale Ordering are also efficient functions used in CNNs.

On a CNN the Max Pooling function also works to prevent overfitting. In the learning model, overfitting is related to the generalization capacity of the model. If the performance of the model with the labeled data is drastically higher than its performance with the unlabeled data, it is assumed that this model does not have the capacity to generalize and, therefore, overfitting.

There are several ways to solve the problem of overfitting. Considering the low performance with the unlabeled data, one of the possible findings of overfitting implies the low variability of examples for the domain mapping. Thus, to solve this problem, increasing the sample space of the labeled data is the solution usually adopted.

Another problem that leads to overfitting concerns the sensitivity in the variation of characteristics. In this case the problem is usually solved by investigating the model's architecture. In the context of the CNN, pooling functions are essential to solve this problem, since these functions aim to make the mapping of invariant characteristics, even if there are variations in the input values. Recent studies demonstrate that Deep Learning models perform well with sensor and temporal data. The work proposed by Zhang, Geiger et al. (2018) shows that DNNs have been proved to be useful in learning from speech data. Sensor data of a smartphone are also kinds of time series with similar characteristics to speech signals.

34.3 Deep Learning applications in computational biology

Computational biology integrates knowledge from different branches of science: biological, computing, mathematics, statistics, chemistry, and physics. At the end of the 1990s, the first biological databases disseminated its content through HTTP protocol (Ouzounis, 2012), which increase in the following years according to the technological evolution of computing science in its various aspects: algorithms, new programming languages, and artificial intelligence methods.

The application of computational methods to process and analyze biological data increased significantly with the emergence of high-throughput sequencing technologies (second- and third-generation sequencing platforms) in the early 21st century (Pearson, 2001) due to their power to decode the genome producing large files with adenine (A), cytosine (C), guanine (G), and thymine (T). Among the main features of these platforms, the increase of data generation and cost reduction by sequenced base (Schuster, 2008), which popularized the sequencing of complete genomes and transcriptomes, in comparison to the first large projects of sequencing: the human genome whose results were released in 2001 by a private and a public consortium (Green, Watson, & Collins, 2015).

With the development of new applications based on DNA decoding techniques, analyses beyond the structural types have gained space, such as evaluation of gene expression that allows the comparison of levels of gene expression under different conditions (stresses) to identify, for example, genes of biotechnological interest; molecular markers to perform a more accurate diagnosis to provide adequate treatment through personalized medicine today supported by data science (Fröhlich et al., 2018); assessment of the interaction of pathogens and their hosts to understand the pathogen invasion system and host defenses (Saha, Sengupta, Chatterjee, Basu, & Nasipuri, 2018); among other studies supported by computational biology analysis/tools.

Conventional data analysis strategies have been challenged due to the rapid increase rate in biological data dimension and acquisition (Kell, 2006). Deep Learning, one branch of Machine Learning methods, has been a very helpful method for finding hidden structures and making predictions using the large quantity of data available (Manyika et al., 2011). In this topic, it is discussed the applications of Deep Learning in omics and biological image processing.

34.3.1 Omics

The evolution of sequencing platforms allowed the growth of another known methods such as genomics, transcriptomics, proteomics, and metabolomics, generating the increase of many applications based on the information obtained through these methods (Zhang et al., 2019). In this section the applications presented will be focused protein structure prediction, gene expression, and protein classification.

DNNs are heavily used in protein structure research (Lyons et al., 2014). The complete prediction, which is performed in three dimensions in space, is challenging and complex. Some studies have tried simpler approaches, such as the prediction of the torsion angles or the secondary structure of a protein.

For instance, Heffernan et al. (2015) used SAE (Sparse Autoencoder) on protein amino acid sequences to solve problems with predictions of torsion angle, secondary structure, and accessible surface area. In a different study, Spencer, Eickholt, and Cheng (2015) used DBN (Deep Belief Networks) to amino acid sequences together with PSSM (Position-Specific Scoring Matrix) and Atchley factors to predict secondary structure of proteins. DNNs have also shown capabilities in gene expression regulation (Leung, Xiong, Lee, & Frey, 2014). Lee and Yoon (2015) used a DBN in the prediction of splice junction and proposed a new DBN method of training named Boosted Contrastive Divergence for imbalanced data and a new term for regularization of sparsity of DNA sequences. This work demonstrated better performance and the ability to show subtle noncanonical signals of splicing.

Chen, Li, Narayan, Subramanian, and Xie (2016) used MLP (multilayer perceptron) to microarray and RNA-seq expression data to infer, from around 1000 landmark genes, the expression of almost 21,000 target genes. When it comes to classification of proteins, Asgari and Mofrad (2015) used the Skip-Gram model, a very known natural language processing method, an MLP variant, and demonstrated that it could learn a distributed representation of biological sequences with general use for several omic applications, protein family classification included.

A few studies have used CNNs to resolve biological sequences problems related to gene expression regulation (Denas & Taylor, 2013). Even so, they have first introduced the advantages of using CNNs, showing great promise for research in the future. First, an initial convolution layer can capture patterns of local sequence and can be considered a detector of motif for which PSSMs are solely learned from data, instead of hard coded. Motifs are short DNA sequences, which represent a kind of recurring patterns that may have a biological function (D'haeseleer, 2006). The CNNs' depth makes learning more complex pattern possible and can capture longer motifs, integrate cumulative effects of motifs that can be observed, and, eventually, learn sophisticated regulatory codes (Park & Kellis, 2015).

Moreover, CNNs may be tailored to explore the benefits of learning by joint multitask. By training CNNs to predict closely factors that are related, simultaneously, features with predictive strengths are learned and shared across different tasks more efficiently. For example, Denas and Taylor (2013) used preprocessed data from ChIP-seq into a matrix with two dimensions where the rows represented the profiles of transcription factors for each gene and a CNN with two dimensions that is very similar to its use in the processing of images. ChIP-seq was first described in 2007. ChIP sequencing as well as microRNA sequencing was one of the first methods to make use of the power of massively parallel or next-generation sequencing to significantly advance real-time PCR (polymerase chain reaction) and array-based methods. ChIP-seq is a counting assay that uses only short reads to align to the genome but requires millions of them to provide meaningful data (Robertson et al., 2007). More recent studies have focused on the usage of CNNs with one dimension and biological sequence data. Alipanahi, Delong, Weirauch, and Frey (2015) and Kelley, Snoek, and Rinn (2016) proposed approaches based on CNNs for the prediction of transcription factor binding site and the multitask prediction of 164 cell-specific DNA accessibility, respectively. The two groups showed downstream applications for the identification of disease-associated genetic variants.

RNNs (Recurrent Neural Networks) are expected to be a decent Deep Learning architecture because of the variable lengths of biological sequences. Studies have shown the successful application of RNNs on the prediction of protein structures (Baldi, Brunak, Frasconi, Soda, & Pollastri, 1999), the regulation of gene expression (Park, Min, Choi, & Yoon, 2016), and the classification of proteins (Hochreiter, Heusel, & Obermayer, 2007). In the first studies, Baldi et al. (1999) used BRNN (Bidirectional Recurrent Neural Networks) with Perceptron hidden units in the prediction of protein secondary structure. Knowing the secondary structure of a protein is important in several aspects, for example, for the design of drugs. After that the LSTM (long short-term memory) hidden units improved performance and became widely recognized. Dediu, Hernández-Quiroz, Martín-Vide, and Rosenblueth (2015) used BRNNs and LSTM hidden units and a convolution layer with one dimension to learn representations from sequences of amino acid and classify the subcellular locations of proteins, then the research identify that proteins localized on membrane, among another, which are common and can be target for drugs.

Other studies explored the functional annotation applying Deep Learning. Guan et al. (2018) proposed a method called Stacked Denoising Autoencoder Multilabel Learning. This tool used Denoising Autoencoder Algorithm and text mining techniques for helping the genes multifunction discovery and the completion of the pathways in cancer data. Results showed that the method exceeds the existing classical multilabel algorithms.

34.3.2 Biological image processing

Probably the most successful field of DNNs is in the image analysis and processing. Deep architectures after trained with millions of photographs can detect objects in photos better than humans (He, Zhang, Ren, & Sun, 2015). All the most significant applications for image classification and retrieval, object detection, and semantic segmentation use DNNs.

The CNN (Fig. 34.6) is the most common deep network architecture for image analysis. In general, a CNN performs pattern matching (convolution) and aggregation (pool) operations. At the pixel level the convolution operation scans the image with a certain pattern and calculates the strength of the match for all positions. The pool step determines the

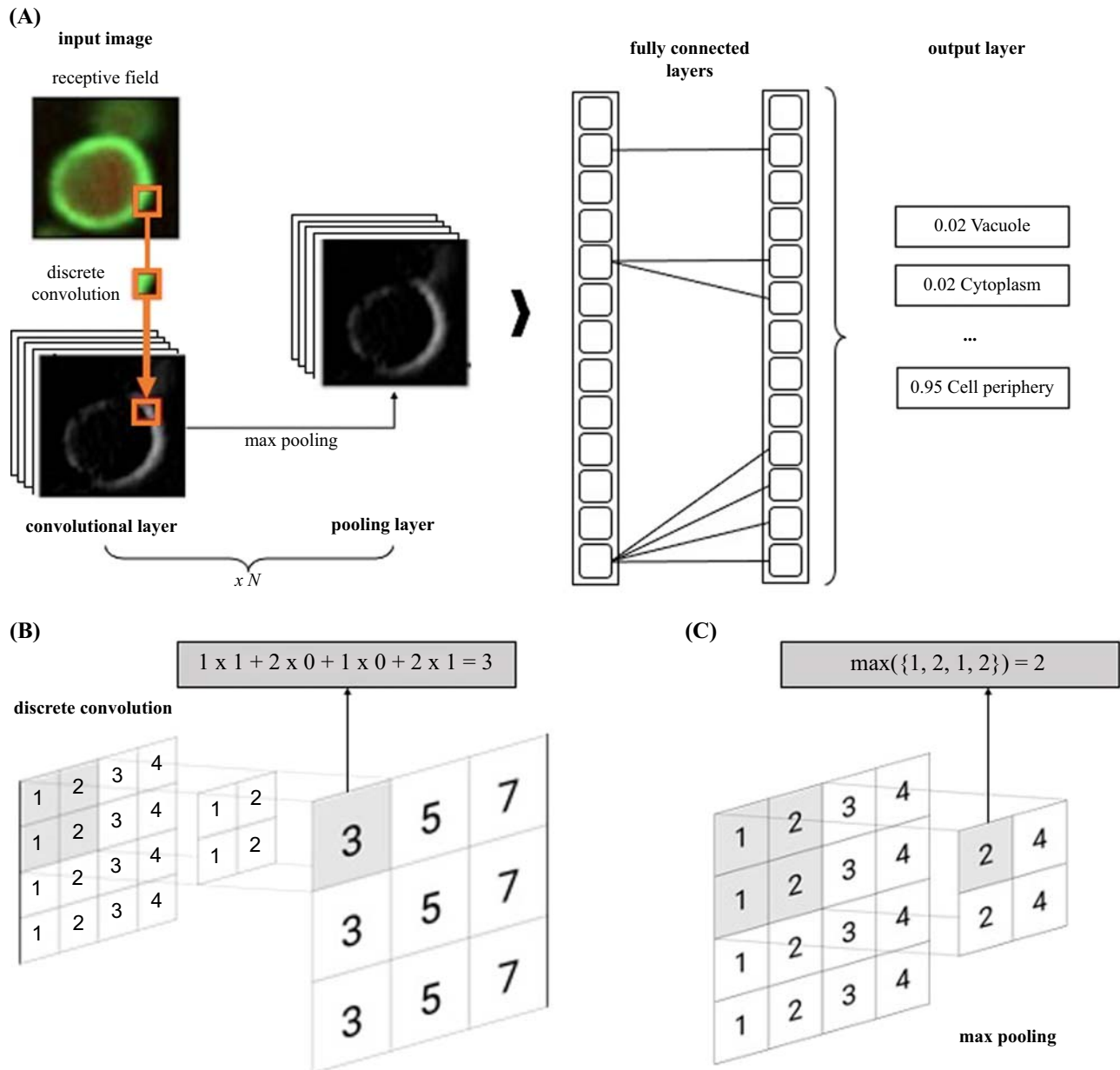


FIGURE 34.6 (A) A convolutional layer with multiple feature maps. (B) The activity of a neuron is obtained by computing a discrete convolution of its receptive field, computing the weighted sum of input neurons, and applying an activation function. (C) Pooling layer summarizes adjacent neurons by computing the maximum or average over their activity, resulting in a smoother representation of feature activities. Adapted from Angermueller, C., Pärnamaa, T., Parts L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, 12, 878. <https://doi.org/10.15252/msb.20156651>.

presence of the pattern in a region, for example, by calculating the maximum pattern matching in smaller patches (Max Pooling), thus aggregating the region's information into a single number. Most network architectures used for image analysis apply a succession of convolution and pooling operations

A convolutional layer consists of several neuron maps, called feature or filter maps, the size of which is equal to the size of the input image (Fig. 34.6A). Two concepts allow to reduce the number of parameters of the model: local connectivity and parameter sharing. First, unlike a fully connected network, each neuron within a feature map is connected only to a local fragment of neurons in the anterior layer, the so-called receptive field. Second, all neurons in a given feature map share the same parameters. Therefore all neurons in a feature map look for the same feature in the previous layer, but in different locations. Maps of different characteristics can, for example, detect edges of different orientation in an image or sequence motifs in a genomic sequence. The activity of a neuron is obtained by calculating a discrete convolution of its receptive field, which is the calculation of the weighted sum of input neurons and the application of an activation function (Fig. 34.6B). In most applications the exact position and frequency of characteristics are irrelevant to the final prediction, such as recognizing objects in an image. Using this assumption, the pool layer summarizes the adjacent neurons by calculating, for example, the maximum or the average over their activity, resulting in a smoother representation of resource activities (image 34.6C). By applying the same grouping operation to small image corrections that are shifted by more than one pixel, the input image is effectively reduced in sampling, further reducing the number of model parameters. One or more fully connected layers can follow the last pool layer (Fig. 34.6A). The hyperparameters of the model, such as the number of convolutional layers, the number of characteristic maps, or the size of the receptive fields, depend on the application and must be rigorously selected from a set of validation data (Angermueller et al., 2016).

Image processing with Deep Learning techniques is also applied in areas such as biomedical image for the clinical treatment of patients (Cerutti et al., 2011). Magnetic resonance imaging (MRI), radiographic imaging, positron emission tomography (PET), and histopathology imaging have been important tools used as input data for Deep Learning algorithms.

Problems of anomaly classification are one of the most studied fields (Plis et al., 2014). Also, in general, tasks related to imaging, segmentation and recognition are often studied in image processing (Woalder, 2017). Some popular high-content screening studies (the quantification of cell biology microscopic images) are also covered by the image processing topic (Woalder, 2017).

DNNs have been used in anomaly classification, segmentation, recognition, and brain decoding. Plis et al. (2014) classified patients with schizophrenia using brain MRIs and DBNs. Woalder (2017) used SAE to identify cell nuclei from images of histopathology.

Most biomedical image processing studies are performed using CNNs. For example, in anomaly classification, Roth et al. (2016) used CNNs to three different CT (computed tomography) dataset images to classify sclerotic metastases, colonic polyps, and lymph nodes. Ciresan, Giusti, Gambardella, and Schmidhuber (2013) used CNNs to detect breast cancer mitoses in histopathology images. Ypsilantis et al. (2015) used PET images esophageal cancer to predict responses to neoadjuvant chemotherapy.

34.3.3 Multiomic data integration

In the last decade, omic data increase exponentially in volume and variety, due to technological progress in DNA sequencing, transcriptomic, proteomic, exome, and other biological fields. Even though these data can elucidate many biological questions individually, their combination and integration can promise new insights about the complex biological systems, illness, tissues, organisms, regulation, and coexpression inside the genome and allowed to understand different behavior, process, and interaction within the biological organisms (Grapov, Fahrman, Wanichthanarak, & Khoomrung, 2018; Li & Ngom, 2015; Li, Wu, & Ngom, 2018).

The multiomic data integration is defined as the incorporation of genomic data from different data omics sources in a meaningful way to provide a more comprehensive analysis of a biological point of interest (Ritchie, Holzinger, Li, Pendergrass, & Kim, 2015). Nowadays, data integration is a significant challenge for the bioinformatics and computational biology, due to the diversity and dimension of those data. Different mathematical and statistical methods, in addition to Machine Learning algorithms, are often used to produce a meaningful and representative integration of omic data (Li & Ngom, 2015; Li et al., 2018; Pinu et al., 2019).

The Deep Learning techniques can help to improve the multiomic data integration and allow more accurate and complete modeling of complex biological systems and processes. Recently, Deep Learning autoencoder techniques showed considerable potential in this branch.

Zhang, Lv et al. (2018) used autoencoder algorithms to integrate multiomic data from neuroblastoma patients and combined with the K-means clustering technique to identify two subtypes with significant survival differences. The application of this technique improved the understanding of the molecular mechanisms and helped the clinicians to make decisions. Chaudhary, Poirion, Lu, and Garmire (2018) used similar techniques to build a sensitive survival model using RNA sequencing, miRNA sequencing, and methylation data from patients with liver cancer. The researchers discovered two optimal subgroups of patients with significant survival differences and functional model fitness.

To explore the implication of Deep Learning in human health, Grapov et al. (2018) presented a literary survey about the challenges and opportunities at a system and biological scale for precision medicine. They said that the DL methods had been shown able to be representing and learning relationships in diverse forms of the omic data and can lead to a transformation in the researches in the precision medicine area.

Maui is a Python (widely used programming language) package developed to simplify the multidata integration by Ronen, Hayat, and Akalin (2019). This computational tool is based on the Bayesian Latent Factor model, the inference of which is done by using ANNs (autoencoder). Using that package, the researchers discovered patterns that describe the variation across the different data modalities, capturing essential aspects of cancer biology, including different gene expression, and mutational profiles, in data from colorectal cancer patients.

In their work, Li, Chen, and Wasserman (2016) proposed a Deep Feature Selection model that takes advantage of the nonlinear Deep Learning structures to make an appropriate subset selection from multiomic data. They applied the model to identify active enhancers and promoters by integrating omic data from the lymphoblastoid cell lines.

34.3.4 Single-cell RNA sequencing

In the last years the single-cell RNA sequencing (scRNA-seq) transformed the genomic and biological science bringing a new way to study heterogeneity in cell populations.

The scRNA-seq can reveal important information about the heterogeneity of complex tissues, cellular states, and profiling the genes expression for a significant number of single cells in parallel. For the analysis of the data produced by scRNA-seq, several methods, pipelines, and algorithms based on Deep Learning have been proposed in the literature (Hwang, Lee, & Bang, 2018; Kolodziejczyk, Kim, Svensson, Marioni, & Teichmann, 2015; Svensson et al., 2017).

scPred is a tool proposed by Alquicira-Hernandez, Sathe, Ji, Nguyen, and Powell (2019) for predicting the cell types, using a combination of unbiased feature selection from a reduced-dimension space and DL algorithms. This tool resolves several problems associated with the identification of individual gene features selection. The authors validated the performance of the pipeline using a dataset to classify tumor versus nontumor epithelial cells in gastric cancer and achieved 99% of accuracy classifying the disease state of individual cells.

The analysis of scRNA-seq data can be obstructed by the noise produced in the amplification and dropout process; for that reason, scalable denoising methods for increasingly large but sparse scRNA-seq data are needed. Eraslan, Simon, Mircea, Mueller, and Theis (2019) proposed an autoencoder network-based method to denoising the scRNA-seq datasets. The pipeline can denoising improves a diverse set of typical scRNA-seq data analyses using simulated and real datasets.

Lopez, Regier, Cole, Jordan, and Yosef (2018) proposed a introduce single-cell variational inference (scVI), a tool for scRNA-seq data processing and analysis. This tool uses nonlinear Deep Learning and variational inference to model library size and batch effect biases, to impute the expression of genes with dropout measurements, and to normalize gene expression matrices, and this allows solver several technological biases. scVI showed a robust performance across many tasks capturing the representation of the appropriate sources of variability for these data (Way & Greene, 2018).

scGen is a tool proposed by Lotfollahi, Wolf, and Theis (2019) to predict the perturbation and infection response of cell types, studies, and species. The tool learns cell type and species-specific, responses signifying that it captures features that identify responding from nonresponding genes and cells. ScGen can be a crucial tool to design in silico screening of perturbation response in the context of disease and drug treatment.

34.3.5 Pharmacogenomics

The goal of pharmacogenomics is to study how the genome-wide variations affect the patterns of pharmacokinetics and pharmacodynamics of individuals. Deep Learning is present in this field in experiments that permit the discovery and develops new drugs using omic data (Relling & Evans, 2015).

The Aliper et al. (2016) team developed and trained a DL model to predict the pharmacological properties of drugs and the repurposing using a large transcriptional response dataset. This work was a demonstration of principles for applying Deep Learning to drug discovery and development.

DL-ADR is algorithm that can classify various SNPs (single-nucleotide polymorphisms) to the corresponding adverse reactions using the Generative Stochastic Networks model. The method allows exploring the complex association between genomic variations and multiple events in pharmacogenomic studies (Liang, Huang, Zeng, & Zhang, 2016).

34.3.6 Modeling biological data in a Deep Neural Network

Most applications in omic sciences can be described within the canonical Machine Learning workflow, which involves four steps: data cleaning, preprocessing, feature extraction, model adjustment, and evaluation (Fig. 34.7A). A supervised Machine Learning model aims to learn a function $f(x) = y$ from a list of training pairs $(x_1, y_1), (x_2, y_2)$, for which data are recorded (Fig. 34.7B). For example, a typical application in biology is to predict the viability of a cancer cell line when exposed to a chosen drug (Eduati et al., 2015; Menden et al., 2013). The input characteristics (x) capture the somatic variants in the cell line sequence, the chemical composition of the drug and its concentration which, together with the measured viability (output label y), can be used to train a support vector machine (SVM), a random forest (RF) classifier, or other related methods (functional relationship f). Given a new cell line (unlabeled data sample x^*) in the future, the learnt function is able to predict its survival (output label y^*) by calculating $f(x^*)$. Methods such as regression (where y is a real number) and classification (where y is a categorical class label) can be viewed in this way. On the other hand, the unsupervised Machine Learning approach aims to discover patterns in the data samples (x), without the need for labeled output data (y). Methods such as clustering, principal component analysis, and outlier detection are typical examples of unsupervised models applied to biological data. The inputs (x), calculated from the raw data, represent what the model “sees in the world” and its choice is highly specific to the problem (Fig. 34.7C). Obtaining the most informative characteristics is essential for the model performance, but the process can be labor intensive and requires domain knowledge. This bottleneck is especially limiting for large data (Angermueller et al., 2016).

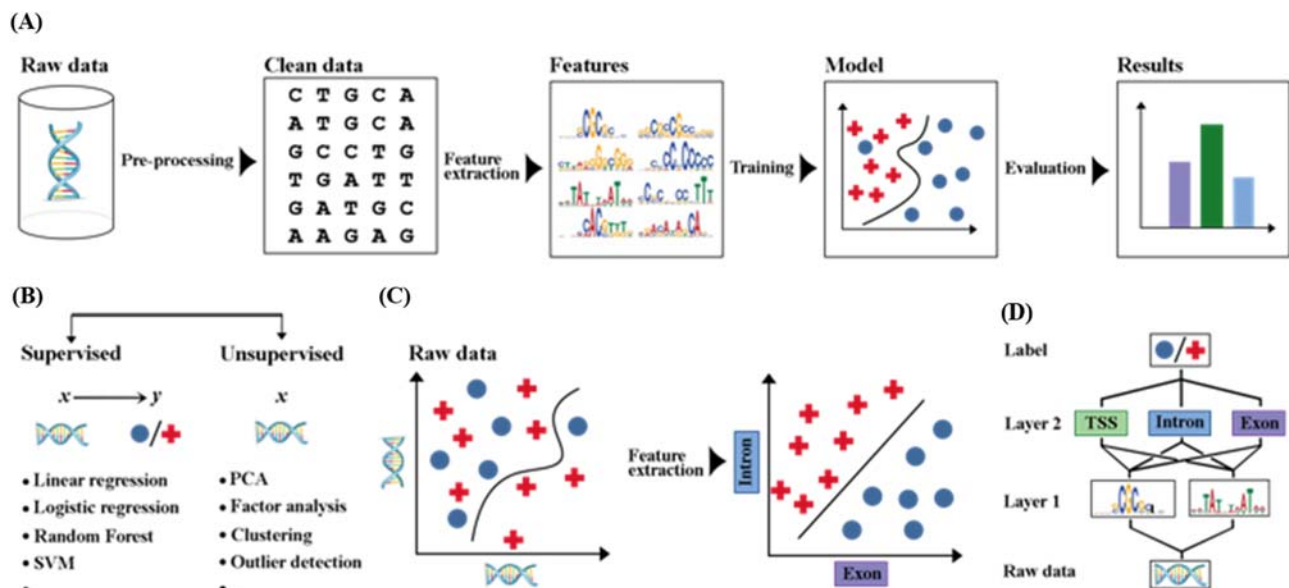


FIGURE 34.7 Machine Learning and representation learning. (A) The four stages of the classic Machine Learning workflow: data preprocessing, feature extraction, learning model, and model evaluation. (B) In supervised learning the input characteristics x are related to the output label y , while unsupervised approach learns factors about x with no observed labels. (C) The raw input data are generally high in dimension and related to the corresponding label in a complex way, which makes it difficult for many classic Machine Learning algorithms (left graph). On the other hand, when extracting higher level characteristics with Deep Learning, it is possible to better distinguish the classes (right graph—correct). (D) Deep networks use a hierarchical layered structure to learn representations of increasingly abstract features from raw data. Adapted from Angermueller, C., Pärnamäa, T., Parts L., & Stegle, O. (2016). *Deep learning for computational biology*. *Molecular Systems Biology*, 12, 878. <https://doi.org/10.15252/msb.20156651>.

A major recent advance in Machine Learning field is the automation of this critical stage, where the model learns an adequate representation of the data through deep ANNs (Fig. 34.7D). In general, a DNN takes the raw data to the lowest layer (input) and transforms them into increasingly abstract representations, successively combining the outputs of the previous layer, in a data-driven way, encapsulating highly complex functions in the process. Deep Learning is nowadays one of the most researched fields in Machine Learning and has been shown to improve performance in image and speech recognition applications, natural language processing, and, more recently, in computational biology.

The potential use for Deep Learning in high-performance biology is immense. It is possible to better explore larger and larger datasets (DNA sequencing, RNA measurements, flow cytometry or automated microscopy) by training networks with multiple layers that capture their internal structure (Fig. 34.7C and D). When learned DNNs discover high-level features, they improve performance over traditional models, increase interpretability, and provide additional understanding of the structure of biological data.

34.3.6.1 Deep Learning for regulatory genomics

Conventional approaches used in regulatory genomics relate sequence variation to changes in molecular features. One way is to leverage variation between genetically diverse individuals to map the quantitative trait loci. This technique is applied to identify regulatory variants that affect gene expression levels (Montgomery et al., 2010; Pickrell et al., 2010), DNA methylation (Bell et al., 2011; Gibbs et al., 2010), histone marks (Grubert et al., 2015; Waszak et al., 2015), and proteome variation (Albert, Treusch, Shockley, Bloom, & Kruglyak, 2014; Battle et al., 2015; Parts et al., 2014) (Fig. 34.8A). However, any mapping approach is intrinsically limited to the variation present in the training

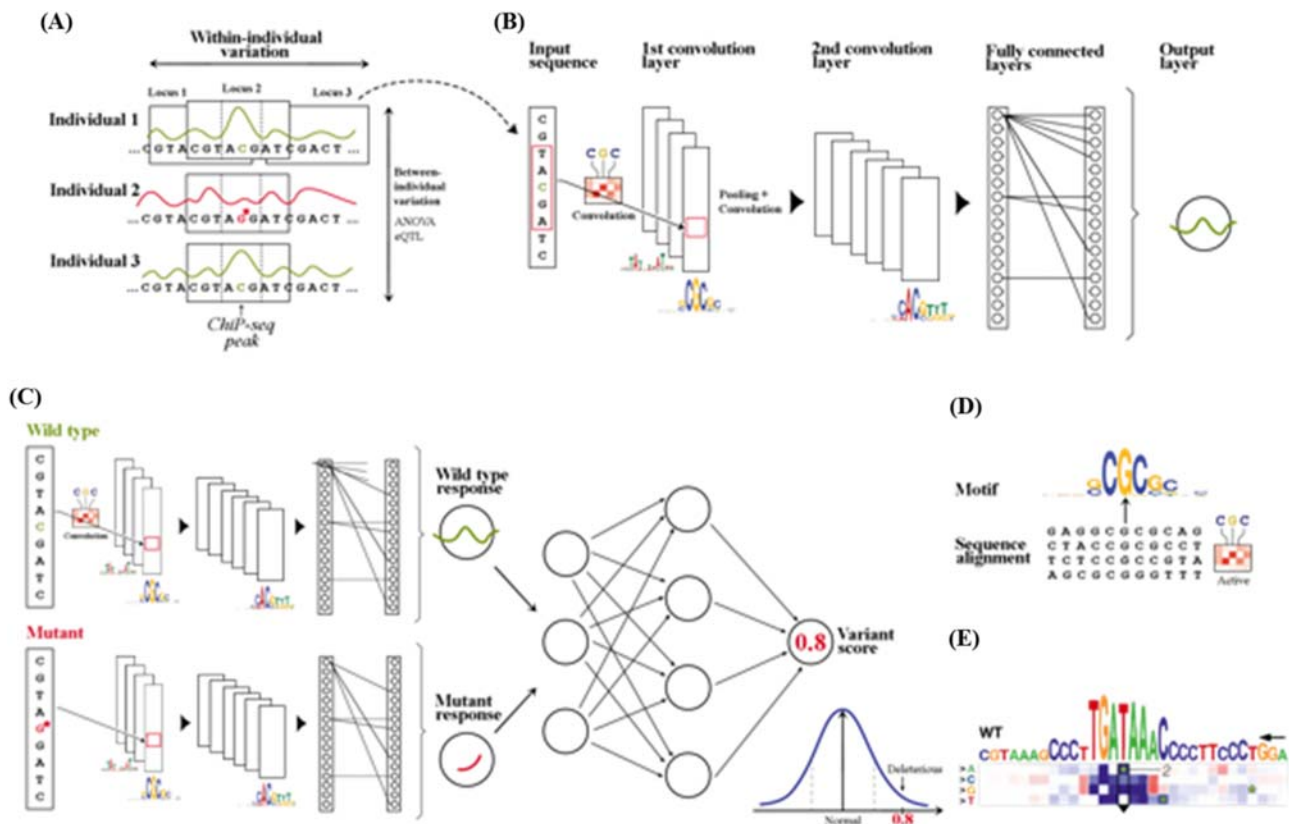


FIGURE 34.8 Using Deep Neural Networks for predicting molecular traits from DNA sequence. (A) DNA sequence and the molecular response variable along the genome for three individuals. Conventional versus Deep Learning approaches. (B) One-dimensional Convolutional Neural Network for predicting a molecular trait from the raw DNA sequence in a window. (C) Response variable predicted by the neural network for a wild-type and mutant sequence is used as input to an additional neural network that predicts a variant score and allows to discriminate normal from deleterious variants. (D) Visualization of a convolutional filter by aligning genetic sequences that maximally activate the filter and creating a sequence motif. (E) Mutation map of a sequence window. Rows correspond to the four possible base pair substitutions, columns to sequence positions. Adapted from Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). *Deep learning for computational biology*. *Molecular Systems Biology*, 12, 878. <https://doi.org/10.15252/msb.20156651>.

population. Thus studying the effects of rare mutations, in particular, requires datasets with very large sample sizes. An alternative is to train models that use variation between regions within a genome (Fig. 34.8A). By dividing the sequence into windows focused on the trait of interest, it is possible to create tens of thousands of training examples for most molecular characteristics, even when using a single individual. Even with large datasets, predicting molecular traits from DNA sequence is challenging due to the multiple layers of abstraction between the effect of individual DNA variants and the trait of interest, as well as the dependence on the molecular traits in a wide sequence context and interactions with distal regulatory elements.

The advantage of using DNNs in this context is twofold. First, classical Machine Learning methods cannot operate directly on the sequence and therefore require predefined features that can be extracted from the sequence based on prior knowledge (e.g., the presence or absence of single-nucleotide variants, k-mer frequencies, motif occurrences, conservation, known regulatory variants, or structural elements). DNNs can help bypass this manual extraction of features by learning directly from data. Second, because of their representational richness, deep networks can capture nonlinear dependencies on the sequence and interaction effects and span broader sequence context at multiple genomic scales (Angermueller et al., 2016).

DNNs have been successfully applied to predict splicing activity (Leung et al., 2014; Xiong et al., 2015), specificities of DNA- and RNA-binding proteins (Alipanahi et al., 2015), or epigenetic marks and to study the effect of DNA sequence changes (Kelley et al., 2016; Zhou & Troyanskaya, 2015).

34.4 Deep Learning applications in agricultural sciences

Several Machine Learning approaches have been applied in agricultural sciences, for example, in crop management, livestock management, water management, and soil management. In crop management, applications can be divided into yield prediction, disease detection, crop quality, weed detection, and species recognition. In animal handling, Machine Learning is already applied to animal welfare and livestock production (Liakos, Busato, Moshou, Pearson, & Bochtis, 2018). In turn, specific Deep Learning techniques have also been applied to problems such as weed identification, land cover classification, plant recognition, fruit counting, crop type classification, prediction of future parameters in corn production, soil humidity in the field, and climatic conditions (Kamilaris & Prenafeta-Boldú, 2018b). Many projects using computer vision techniques addressed to agriculture have been implemented. Research has highlighted the possibilities of applying computer vision systems in agriculture fields such as animal behavior analysis, precision agriculture and machine orientation, silviculture, measurement, and growth plantation analysis (Brosnan & Sun, 2002). Other works have been carried out recently on the identification and classification of weeds (Ferreira, 2017).

In the context of forestry, activities such as grading and classification of wood from their defects are considered tiring and repetitive, and when performed by humans, in a nonautomated way, usually result in unreliable outcomes. According to Kline, Surak, and Araman (2018), the hit rate of human graders is 48% in the sawing line, demonstrating unsatisfying results. Therefore automated image classification methods were developed aiming at decreasing this problem, gaining promising results with hit rates around 90.5% (Rall, 2010) and 96.9% (Almeida, 2014). Processing method and image classification have been suggested for a long time, and many of them use techniques of artificial intelligence (de Almeida, Gomes, de Almeida, & Ballarin, 2018). The classification of forest species is another essential process in proper forest management and forest control. After logging, many species characteristics are lost, and the identification turns out to be an even harder task. It is, then, necessary a wood anatomy analysis, usually carried out by specialists who know very well the cellular structures present in each species. However, this methodology implies in poorly automated techniques, which makes the task time-consuming and error prone. Those factors hinder the control and decision-making by environmental organizations. The use of computer vision techniques is an alternative to automated recognition since it allows the construction of intelligent models that, from the images, are capable of detecting features and performing the final classification. The use of CNN, a Deep Learning technique, has shown to be efficient for this type of application (Oliveira, 2018).

When we consider the Animal Science area, for example, the beef production chain has sought to generate products that meet the requirements of the final link in this chain, that is, beef consumers. In this context, one of the approaches used is the evaluation of carcass quality, which aims to estimate the characteristics of the meat produced. The use of methods to assess carcass quality after slaughtering animals is of little benefit, and it is recommended to use methods applied at the time of purchase or to separate animals into slaughter lots. The ultrasound technique allows the evaluation of carcass characteristics through a noninvasive procedure, without leaving harmful residues in the meat of the animals (Cardoso, 2013; Yokoo et al., 2009). Among the characteristics of bovine carcass measured by ultrasonography, the rib eye area, the thickness of subcutaneous fat, and the thickness of fat on the animal rump can be analyzed (Yokoo et al.,

2015). Despite its benefits, the accuracy of ultrasound measurements on carcass characteristics has generated a wide range of results, which is attributed mainly to the different equipment configurations and the subjectivity inherent to the technician responsible for carrying out the evaluation (Greiner, 2012). Automated approaches, based on CNNs, have been used to estimate the thickness of fat on the animal rump from images obtained ultrasonographically (Bragamonte, Camargo, Cardoso, Yokoo, & Cardoso, 2018).

34.4.1 Example of Deep Learning applied to agriculture

Various computational approaches are currently used for detecting plant diseases and most common are ANNs and SVMs. They are combined with different image preprocessing methods to achieve better feature extraction. The method described in Sladojevic, Arsenovic, Anderla, Culibrk, and Stefanovic (2016) is a novel plant disease recognition model, based on leaf image and using a deep CNN trained and fine-tuned to fit accurately to the database of a plant's leaves that was gathered independently for diverse plant diseases. The advance and novelty of the developed model lie in its simplicity: healthy leaves and background images are in line with other classes, enabling the model to distinguish between diseased leaves and healthy ones or from the environment by using deep CNN. An example of Deep Learning architecture used in this example is shown in Fig. 34.9, which illustrates CaffeNet (Jia et al., 2014), a classical type of CNN, combining convolutional and fully connected (dense) layers.

As Fig. 34.9 shows, several convolutions are carried out in some layers of the network, creating different representations of the learning dataset, starting with the most general ones in the first and largest layers, becoming more specific in the deeper layers. The convolutional layers act as features extractors of the input images, the dimensionality of which is then reduced by the pooling layers. Convolutional layers encode multiple low-level features into more discriminating features, so that they recognize the context spatially. They can be understood as filter banks that transform one input image into another, highlighting specific patterns. Fully connected layers, located in many cases close to the model's output, act as classifiers that exploit the high-level features learned to classify input images into predefined classes or to make numerical predictions. They use a matrix as an input and produce another matrix as an output (Sladojevic et al., 2016).

An example of visualization of leaf images after each processing step of CaffeNet CNN applied to a plant diseases identification problem is shown in Fig. 34.10. It is observed that after each processing step, the image-specific elements that indicate a possible disease become more evident, especially in the final stage (Pool5). The output images are labeled with the name of the corresponding layer in the lower right corner of each image.

From historical research statistics or recent research results, it is possible to infer that Deep Learning techniques can be well applied in agricultural science to solve several problems that have been worrying farmers and scientists for a long time. With the help of new techniques and theories, these new approaches end up surpassing traditional methods in several aspects (Zhu et al., 2018).

34.4.2 Convolutional Neural Networks in agriculture

Smart agriculture (Tyagi, 2016) is essential to face the diverse agricultural universe challenges, such as productivity, environmental impact, food security, and sustainability (Gebbers & Adamchuk, 2010). The steady growth of the global

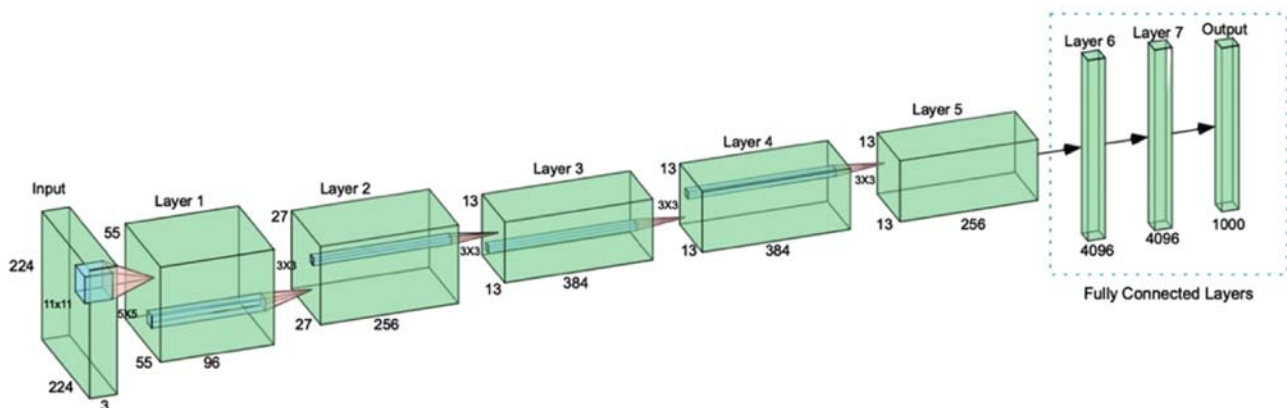


FIGURE 34.9 CaffeNet, an example of CNN architecture. CNN, Convolutional Neural Network.

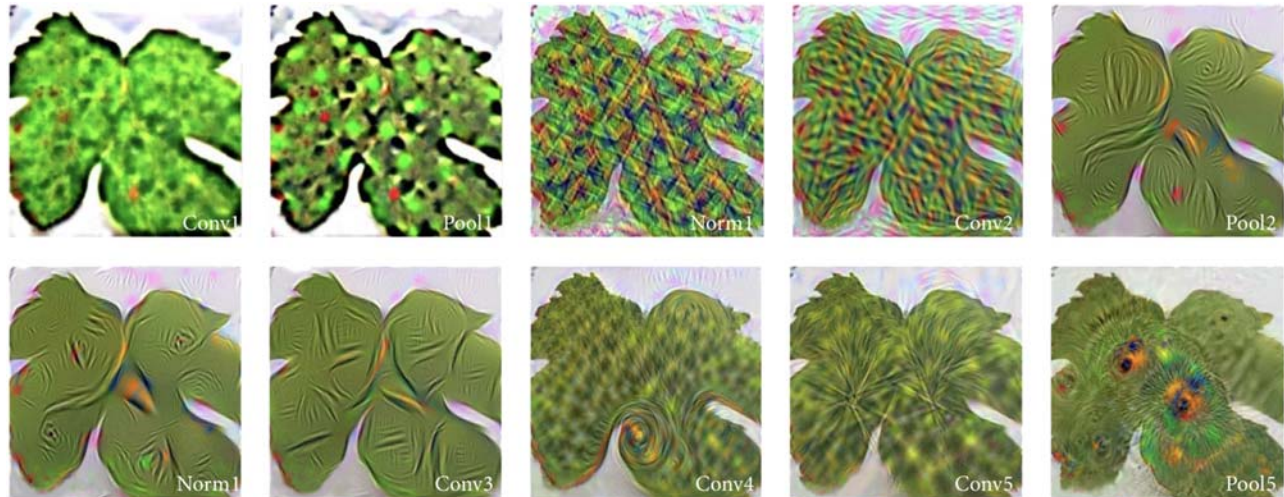


FIGURE 34.10 Visualization of the output layer images after each CaffeNet CNN processing step (convolution, pooling, normalization) in a plant disease identification problem based on leaf images. CNN, Convolutional Neural Network. Based on Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). *Deep neural networks based recognition of plant diseases by leaf image classification*. Computational Intelligence and Neuroscience, 2016, 3289801, 11 pages. Hindawi Publishing Corporation. <https://doi.org/10.1155/2016/3289801>.

population (Kitzes et al., 2008) demands a substantial increase in food production (FAO, 2009). On the other hand, the protection of forests and natural ecosystems cannot be ignored through sustainable agricultural procedures. Food needs to maintain a high nutritional value, while food security needs to be guaranteed worldwide (Carvalho, 2006).

To face these challenges, complex, multivariate, and unpredictable agricultural ecosystems need to be better understood. This knowledge can be achieved by continuously monitoring, measuring, and analyzing various aspects and physical phenomena. The implementation of new Information and Communication Technologies (ICT) in agricultural management of small-scale cultivation and harvesting and the observation of ecosystems on a larger scale will facilitate this task, improving management and political decision-making from the context, situation, and awareness location (Kamilaris & Prenafeta-Boldú, 2018a).

Emerging ICT technologies relevant to understanding agricultural ecosystems include remote sensing (Bastiaanssen, Molden, & Makin, 2000), Internet of Things (IoT) (Weber & Weber, 2010), Cloud Computing (Hashem et al., 2015), and Big Data analysis (Chi et al., 2016; Kamilaris & Prenafeta-Boldú, 2017). Remote sensing, using satellites, airplanes, and unmanned aerial vehicles (UAV, or drones), provides an instant large-scale view of the agricultural environment. These technologies offer several advantages when applied to agriculture, being a well-known and nondestructive method for collecting information on soil characteristics. Remote sensing data can be obtained systematically over large geographic areas, including areas inaccessible to human exploration. The IoT uses advanced sensor technology to measure various parameters in the field, while cloud computing is employed for collecting, storing, preprocessing, and modeling large volumes of data from various heterogeneous sources. Finally, Big Data analysis is employed in combination with cloud computing for real-time, large-scale analysis of data stored in the cloud (Kamilaris, Gao, Prenafeta-Boldú, & Ali, 2016; Waga & Rabah, 2014). These four technologies (remote sensing, IoT, cloud computing, and big data analysis) can create new applications and services to improve agricultural productivity and increase food security, such as understanding climate conditions and changes better.

Much of the amount of data collected by remote sensing and IoT consists of images. These data can provide a complete view of agricultural farmlands, and the analysis of these images can help in a variety of important issues (Liaghat & Balasundram, 2010; Ozdogan, Yang, Allez, & Cervantes, 2010). Therefore image analysis is an essential area of research in the agricultural domain, and intelligent analysis techniques are applied for image identification, classification, and anomaly detection in different agricultural applications (Saxena & Armstrong, 2014; Teke, Deveci, Haliloğlu, Gürbüz, & Sakarya, 2013). Among these, the most common detection method is based on satellite, using multispectral and hyperspectral images. Synthetic aperture radar (SAR), thermal and near-infrared cameras (thermal and near-infrared cameras) are used to a lesser extent (Ishimwe, Abutaleb, & Ahmed, 2014), while optical and radiographic images are already commonly applied in the classification of packaged fruits and foods (Saxena & Armstrong, 2014).

The most commonly used techniques for image analysis are based on Machine Learning algorithms, such as K-means, SVMs, and ANNs, as well as filtering based on wavelets, vegetation indexes such as the Normalized Difference Vegetation

Index and Regression Analysis (Saxena & Armstrong, 2014). Also, Deep Learning techniques (LeCun et al., 2015; Schmidhuber, 2015) have been widely applied to image analysis (computer vision) in agricultural sciences. This section presents some problems in agriculture that have successfully employed two particular classes of Deep Learning method. One is called CNNs, defined as a deep ANN (Szegedy et al., 2015) and the second is the RNNs (Mandic & Chambers, 2001) that can take advantage of neural networks for end-to-end classification of a time series. CNN is probably the most popular and widely used technique in agricultural research today, especially in problems related to image analysis. This is due to its high potential to address challenges in agriculture-related to computer vision.

The comparative study reported by Kamilaris and Prenafeta-Boldú (2018a) lists several relevant related works, indicating the problems they addressed, the agricultural area involved, data sources used, general precision achieved, and implementation details based on CNN, as well as comparisons with other techniques, whenever available. Twelve areas have been identified in total, the most popular being: disease detection on plants and leaves, land cover classification, plant recognition, fruit counting, and weed identification. Most of these works address image classification and cataloging of interest areas, including obstacle detection (Christiansen, Nielsen, Steen, Jørgensen, & Karstoft, 2016; Steen, Christiansen, Karstoft, & Jørgensen, 2016) and fruit counting (Rahnemoonfar & Sheppard, 2017; Sa et al., 2016), while other studies focus on making predictions of future benefits, such as the corn production (Kuwata & Shibasaki, 2015) and the soil moisture content in the field (Song et al., 2016). From another perspective, most researches focus on productions, while few consider land cover and livestock issues. Several performance metrics were used by the authors, with the percentage of correct predictions (Classification Accuracy, CA) being the most used in validating or testing the dataset. Other works used Root-Mean-Square Error (RMSE), F1 Score, Quality Measure (QM) (Douarre, Schielein, Frindel, Gerth, & Rousseau, 2016), Residual Functional Capacity (RFC) (Chen et al., 2017), and LC (Reyes, Caicedo, & Camargo, 2015). The majority of those studies employed CA, which is generally high (i.e., over 90%), indicating the successful application of CNN to various agricultural problems. Table 34.1 shows a comparison between these methods that applied CNNs to problems in agriculture.

The comprehensive analysis of Kamilaris and Prenafeta-Boldú (2018a) showed that CNN offers superior performance in terms of accuracy in the vast majority of applications presented, based on the performance metrics employed by the authors, with the Gaussian Mixture Model (GMM) being a technique with comparable performance in some cases (Reyes et al., 2015; Santoni et al., 2015). In most of the problems from agricultural areas, satisfactory accuracy was observed, especially compared to other techniques applied to solve the same problem. These results show the successful application of CNN in a variety of agricultural domains. In particular, the areas of disease detection in plants and leaves, plant recognition, classification of soil covering, fruit count, and weed identification belong to the categories in which the highest accuracy was observed. Although CNN has been associated with computer vision and image analysis, two related works have been found in which CNN-based models are trained based on field sensory data (Kuwata & Shibasaki, 2015) and a combination of static and dynamic environments (Song et al., 2016). In both cases the performance (i.e., RMSE) was better than other techniques under consideration.

When comparing performance in terms of accuracy and exactness, the same experimental conditions and metrics must exist for evaluating datasets and performance (when comparing CNN with other techniques), as well as architectures and models parameters (when comparing studies using CNN).

The study of Kamilaris and Prenafeta-Boldú (2018a) showed only 12 types of problems related to agriculture, in which CNN technique was successfully used. It would be interesting to comprehend how CNN behaves in other sectors of agriculture, such as crop phenology, seed identification, nitrogen content in soil and leaves, irrigation, water stress detection in plants, water erosion assessment, pest detection and herbicides use, contaminants identification, diseases or defects in food, damage to crop hail and monitoring of greenhouses. Several of these research areas employ data analysis techniques with similar concepts and performance comparable to CNN, such as linear and logistic regression, SVM, K-nearest neighbor (KNN), K-means clustering, wavelet-based filtering, Fourier transform, and would be worth examining CNN's applicability to these problems. Another possible area of application for CNNs would be using images, employing drones, to monitor the effectiveness of the sowing process and increase the quality of production, for example, of wine, harvesting the grapes at the right time to obtain the best levels maturity. It can also be applied in monitoring animals and their movements to consider their well-being and identify possible diseases, in addition to many other scenarios where computer vision is involved.

As observed, the cited works cited used standard CNN architectures that constitute only a specific and particular category of Deep Learning models. Advanced and sophisticated models such as RNNs (Mandic & Chambers, 2001) have been also tested in agricultural sciences tasks. These architectures tend to exhibit a dynamic temporal behavior and are capable of remembering, but also forgetting after some time or when necessary. An application example could be to estimate the growth of plants, trees, or even animals based on previous consecutive observations, to predict their yield,

TABLE 34.1 Applications of Deep Learning in agriculture.

No.	Agricultural area	Problem description	Data used	Precision	DL model used	DL framework used	Comparison with other techniques	Ref.
1	Leaf disease detection	13 different types of plant diseases, plus healthy leaves	Authors-created database containing 4483 images	96.30% (CA)	CaffeNet	Caffe	Better results than SVM (no more details)	Sladojevic et al. (2016)
2	Plant disease detection	Identify 14 crop species and 26 diseases	PlantVillage public dataset of 54,306 images of diseased and healthy plant leaves	0.9935 (F1)	AlexNet, GoogleNet	Caffe	Benchmarks with approaches using hand-engineered features	Mohanty et al. (2016)
3	Plant disease detection	Classify banana leaf diseases	Dataset of 3700 images of banana diseases obtained from the PlantVillage dataset	96%+ (CA), 0.968 (F1)	LeNet	Deeplearning4j	Methods using hand-crafted features do not generalize well	Amara et al. (2017)
4	Land cover classification	Identify 13 different land cover classes in KSC and nine different classes in Pavia	A mixed vegetation site over KSC, FL, United States, and an urban site over the city of Pavia, Italy	98.00% (CA)	Hybrid of PCA, autoencoder, and logistic regression	Developed by the authors	1% more precise than RBF-SVM	Chen et al. (2014)
5	Land cover classification	Identify 21 land use classes containing a variety of spatial patterns	UC Merced Land Use dataset	93.48% (CA)	Author defined	Theano	UFL (82%–90%) and SIFT (85%)	Luus et al. (2015)
6	Land cover classification	Extract information about cultivated land	Images from UAV at the areas Pengzhou County and Guanghan County Sichuan Province, China	88%–91% (CA)	Author defined	N/A	N/A	Lu et al. (2017)
7	Crop type classification	Classification of crops wheat, maize, soybean sunflower, and sugar beet	19 multitemporal scenes acquired by Landsat-8 and Sentinel-1A R satellites from a test site in Ukraine	94.60% (CA)	Author defined	Developed by the authors	Multilayer perceptron (92.7%), random forests (88%)	Kussul et al. (2017)

8	Plant recognition	Recognize seven views of different plants: entire plant, branch, flower, fruit, leaf, stem and scans	LifeCLEF 2015 plant dataset, which has 91 759 images distributed in 13,887 plant observations	48.60% (LC)	AlexNet	Caffe	20% worse than local descriptors to represent images and KNN, dense SIFT, and a GMM	Reyes et al. (2015)
9	Plant recognition	Recognize 44 different plant species	MK Leaf Dataset which consists of 44 classes, collected at the Royal Botanic Gardens, Kew, England	99.60% (CA)	AlexNet	Caffe	SVM (95.1%), ANN (58%)	Lee et al. (2015)
10	Plant recognition	Identify plants from leaf vein patterns of white, soya, and red beans	866 leaf images provided by INTA Argentina. Dataset divided into three classes: 422 images correspond to soybean leaves, 272 to red bean leaves, and 172 to white bean leaves	96.90% (CA)	Author defined	Pylearn2	PDA (95.1%), SVM and RF slightly worse	Grinblat et al. (2016)
11	Segmentation of root and soil	Identify roots from soils	Soil images coming from X-ray tomography	QM = 0.23 (simulation) QM = 0.57 (real roots)	Author-defined CNN with SVM for classification	MatConvNet	N/A	Douarre et al. (2016)
12	Crop yield estimation	Estimate maize yield at county level in the United States	Maize yields from 2001 to 2010 in Illinois, United States, downloaded from Climate Research Unit (CRU), plus MODIS Enhanced Vegetation Index	RMSE = 6.298	Author defined	Caffe	SVR RMSE = 8.204	Kuwata and Shibasaki (2015)
13	Fruit counting	Predict number of tomatoes in images	24,000 synthetic images produced by the authors	91% (RFC) 1.16 (RMSE) on real images, 93% (RFC) 2.52 (RMSE) on synthetic images	Inception-ResNet	TensorFlow	ABT (66.16%), RMSE = 13.56	Rahnemoonfar and Sheppard (2017)
14	Fruit counting	Map from input images of apples and oranges to total fruit counts	711,280 Å ~ 960 orange images (day time) and 211,920 Å ~ 1200 apple images (night time)	0.968 (RFC), 13.8 (L2) for oranges 0.913 (RFC), 10.5 (L2) for apple	CNN (blob detection and counting) + linear regression	Caffe	Best texture-based regression model (ratio of 0.682)	Chen et al. (2017)

(Continued)

TABLE 34.1 (Continued)

No.	Agricultural area	Problem description	Data used	Precision	DL model used	DL framework used	Comparison with other techniques	Ref.
15	Fruit counting	Fruit detection in orchards, including mangoes, almonds, and apples	Images of three fruit varieties: apples (726), almonds (385) and mangoes (1154), captured at orchards in Victoria and Queensland, Australia	F1 Scores of 0.904 (apples) 0.908 (mango) 0.775 (almonds)	Faster Region-based CNN with VGG16 model	Caffe	ZF network (F1 Scores of 0.892, 0.876, and 0.726 for the apples, mangoes, and almonds, respectively)	Bargoti and Underwood (2017)
16	Fruit counting	Detection of sweet pepper and rock melon fruits	122 images obtained from two modalities: color (RGB) and NIR	0.838 (F1)	Faster region-based CNN with VGG16 model	Caffe	Conditional Random Field to model color and visual texture features (F1 = 0.807)	Sa et al. (2016)
17	Obstacle detection	Identify ISO barrel-shaped obstacles in row crops and grass mowing	A total of 437 images from authors' experiments and recordings	99.9% in row crops and 90.8% in grass mowing (CA)	AlexNet	Caffe	N/A	Steen et al. (2016)
18	Obstacle detection	Detect obstacles that are distant, heavily occluded, and unknown	Background data of 48 images and test data of 48 images from annotations of humans, houses, barrels, wells, and mannequins	0.72 (F1)	AlexNet and VGG	Caffe	Local decorrelated channel features (F1 = 0.113)	Christiansen et al. (2016)
19	Identification of weeds	Classify 91 weed seed types	Dataset of 3980 images containing 91 types of weed seeds	90.96% (CA)	PCANet + LMC classifiers	Developed by the authors	Better results than feature extraction techniques (no details)	Xinshao and Cheng (2015)
20	Identification of weeds	Classify weed from crop species based on 22 different species in total	Dataset of 10,413 images, taken mainly from BBCH 12–16 containing 22 weed and crop species at early growth stages	86.20% (CA)	Variation of VGG16	Theano-based Lasagne library for Python	Local shape and color features (42.5% and 12.2%, respectively)	Dyrmann et al. (2016)

21	Identification of weeds	Identify thistle in winter wheat and spring barley images	A total of 4500 images from 10, 20, 30, and 50 m of altitude captured by a Canon PowerShot G15 camera	97.00% (CA)	DenseNet	Caffe	Color feature–based Thistle Tool (95%)	Sørensen et al. (2017)
22	Prediction of soil moisture content	Predict the soil moisture content over an irrigated cornfield	Soil data collected from an irrigated cornfield (an area of 22 km ²) in the Zhangye oasis, Northwest China	RMSE = 6.77	DBN-MCA	Developed by the authors	MLP-MCA (18% RMSE reduction)	Song et al. (2016)
23	Cattle race classification	Practical and accurate cattle identification from five different races	A total of 1300 images created by the authors	93.76% (CA)	GLCM-CNN	Deep Learning Matlab Toolbox	Deep Learning Matlab Toolbox	Santoni, Sensuse, Arymurthy, and Fanany (2015)

ABT, Area-based technique; *CA*, Classification Accuracy; *CNN*, Convolutional Neural Network; *DBN-MCA*, Deep Belief Network–based macroscopic cellular automata; *GLCM*, Gray Level Cooccurrence Matrix; *GMM*, Gaussian Mixture Model; *KNN*, K-nearest neighbor; *KSC*, Kennedy Space Center; *MK*, MalayaKew; *MLP*, multilayer perceptron; *NIR*, near-infrared; *PDA*, Penalized Discriminant Analysis; *QM*, Quality Measure; *RFC*, Residual Functional Capacity; *RMSE*, Root-Mean-Square Error; *SVM*, support vector machine; *SVR*, support vector regression; *UFL*, unsupervised feature learning; *PCA*, Principal Component Analysis; *DL*, Deep Learning; *RBF*, Radial Basis Function; *SIFT*, Scale Invariant Feature Transform.

Based on Kamilaris, A., & Prenafeta-Boldú, F.X. (2018a). A review of the use of convolutional neural networks in agriculture. *The Journal of Agricultural Science*, 156, 312–322. <https://doi.org/10.1017/S0021859618000436>.

assess their water needs, or prevent disease. Models like these can also find applicability in environmental computer science to understand climate changes and predict climatic conditions and phenomena, estimating the environmental impact of various physical or artificial processes (Kamilaris & Prenafeta-Boldú, 2018a).

34.4.3 Recurrent Neural Network for agricultural classification

Spatial information on agricultural practices plays an essential role in the sustainable development of agronomy, environment, and economy (Buckley & Carney, 2013; Foley et al., 2005). Indeed, the international community, such as the Food and Agriculture Organization, recognized the importance of agricultural practices (Polsot, Speedy, & Kueneman, 2004). In this context, remote sensing satellite images are valuable guidance in understanding the spatial distribution of agricultural cultures. In recent years, many satellites have been launched to acquire high spatial resolution data in various spectral domains. The European Space Agency's (ESA) Sentinel-1 radar and Sentinel-2 optical sensors are suitable for monitoring agricultural areas. However, like all optical sensors, the use of Sentinel-2 data is limited if the cloud layer is large (Drusch et al., 2012). On the other hand, Sentinel-1 is a SAR system that can acquire images in any weather to provide data regardless of weather conditions. SAR data can, for example, distinguish rice from other types of vegetation covering (Le Toan et al., 1997). The ESA Sentinel-1 SAR type sensor allows a precise temporal follow-up of agricultural crop growth (Torres et al., 2012). ESA provides free data that allows obtain exceptional agricultural monitoring for various applications, particularly to provide a detailed spatial agricultural land cover distribution (Ndikumana, Minh, Baghdadi, Courault, & Hossard, 2018). In the Camargue region, in France, agriculture is a significant activity. Among agricultural practices, rice cultivation is the one that stands out the most and plays a crucial role in the development of agriculture systems because rice irrigation allows the leaching of salt and, consequently, the introduction of other species in crop rotation (Mouret, 1988). In this context, the spatial extent of agricultural land cover is primordial.

Remote sensing for classification is usually performed based on supervised Machine Learning approaches (Friedl & Brodley, 1997; Li, Wang, Wang, Hu, & Gong, 2014). Several supervised learning algorithms are available and applicable, each with its strengths and weaknesses (Friedl & Brodley, 1997; Lu & Weng, 2007; Waske & Braun, 2009). The most recent methodological developments are based on approaches of active learning and semisupervised learning, which make use of unlabeled data for training (Gomez-Chova, Camps-Valls, Munoz-Mari, & Calpe, 2008; Li, Bioucas-Dias, & Plaza, 2010; Munoz-Mari, Bovolo, Gomez-Chova, Bruzzone, & Camp-Valls, 2010; Tuia, Volpi, Copa, Kanevski, & Munoz-Mari, 2011); however, the use of these approaches is not yet widespread in agricultural land cover classifications. In practice, for agricultural applications, most remote sensing work is based on traditional algorithms, such as KNN, RF, and SVM (Flamary, Fauvel, Mura, & Valero, 2015; Inglada, Vincent, Arias, & Marais-Sicre, 2016).

Nevertheless, these approaches were not designed to work with time series data and, therefore, ignore their time dependency. On the other hand, the DNNs consider the temporal correlation of the data. By recent advances in Machine Learning, there has been an increasing interest in classifying time series using deep CNNs and RNNs that can take advantage of neural networks for end-to-end classification time series (Ho Tong Minh et al., 2018; Ienco, Gaetano, Dupaquier, & Maurel, 2017; Kamilaris & Prenafeta-Boldú, 2018b). Besides, RNN approaches can work on a pixel-based time series (Ho Tong Minh et al., 2018) once those networks are ideal for this class of classification. Due to their properties, RNNs offer models to explicitly manage dependencies between data, for example, with LSTM (Hochreiter & Schmidhuber, 1996) and Closed Recurrent Unit (Cho et al., 2014), which makes them suitable for mining the Sentinel-1 SAR multitemporal data (Ndikumana et al., 2018).

Ndikumana et al. (2018) in their research used 921 reference plots (July 2017) to collect land cover information. The limit of the reference plots was drawn manually with ArcGIS online. Eleven surface classes observed were chosen: (1) rice, (2) sunflower, (3) lawn, (4) irrigated grassland, (5) durum wheat (winter), (6) alfalfa, (7) tomato, (8) melon, (9) clover, (10) swamps, and (11) vineyard. Fig. 34.11 shows the ground position of the samples and the distribution of the pixel number per class and the number of plots.

Once the wheat cultivation is the only winter crop presence after May and the primary agricultural practices in Camargue are conducted in the summer (from May to September), the +/1B SAR dataset included 25 acquisitions in Terrain Observation by Progressive Scans (TOPS) imaging mode from May to September 2017 (5 months) with a revisiting period of 6 days. First, the master image was chosen, and all images were coregistered, taking into account the TOPS mode, for the master image (Prats-Iraola, Scheiber, Marotti, Wollstadt, & Reigber, 2012). Images with the intensity of five range views (5-range looks) were generated and radiometrically calibrated for range spreading loss, antenna gain, normalized reference area, and constant calibration that depends on the parameters in the Sentinel-1 SAR header. After preprocessing and filtering, all processed images are in the imaging geometries of the master image. In a unified

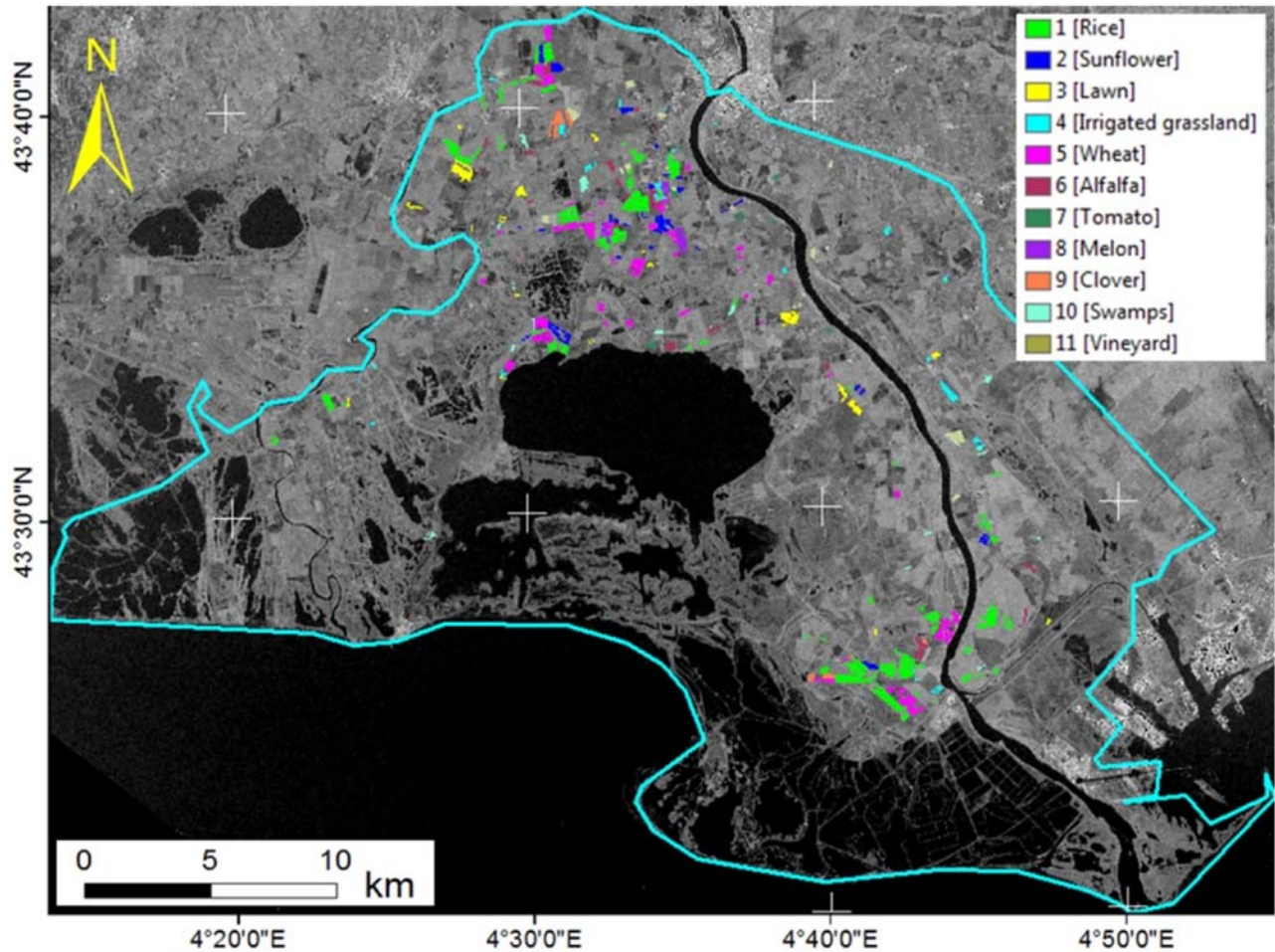


FIGURE 34.11 Camargue study area. Colored polygons represent 921 reference plots location. The study area is limited by the cyan polygon. *Based on Ndikumana, E., Minh, D.H.T., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. Remote Sensing, 10, 1217. doi:10.3390/rs10081217.*

dataset, all data in the image need to be orthorectified into map coordinates. The pixel size of the orthorectified image data is 20 m. After geocoding, all intensity images are transformed to a logarithmic dB scale, normalized to values between 0 and 255 (8 bits), and submitted to the classifiers.

RNNs are well-designed Machine Deep Learning techniques that stand out for their qualities in different domains, such as signal processing, natural language processing, and speech recognition (Linzen, Dupoux, & Goldberg, 2016; Soma, Mori, Sato, Furumai, & Nara, 2015). Unlike CNNs, RNNs manage data dependencies, since the neuron output at time $t - 1$ is used with the next input to feed the neuron itself at time t . A diagram of a typical neural RNN is detailed in Fig. 34.12. Among the different models of RNN, we have LSTM and Gated Recurrent Unit (GRU), which are the best well-known RNN models. The main difference between them is related to the number of parameters to learn. Considering the same size as the hidden state, the LSTM model has more parameters than the GRU unit.

A deep architecture is built in each RNN unit to perform the classification, in a similar way to the CNN structure with several convolutional layers (Bengio, Courville, & Vincent, 2013). The arrangement serves to extract high-level nonlinear time dependencies which are in the remote sensing time series. This structure is similar to both LSTM and GRU. The RNN model follows a new sequence at the input but does not predict by itself. For this, a SoftMax layer (Graves, Mohamed, & Hinton, 2013) is connected to the last recurrent unit to predict the final multiclass. The SoftMax layer has the same number of neurons as the classes to be predicted. Each sample belongs to only a single class, which leads to the choice of SoftMax. This scheme is instantiated for the LSTM and GRU units, creating two different classifiers: a classification scheme based on LSTM and one based on GRU. Fig. 34.13 shows the LSTM-based architecture scheme for each pixel (25 input points, 5 LSTM units, 512 hidden dimensions, and 11 output classes).

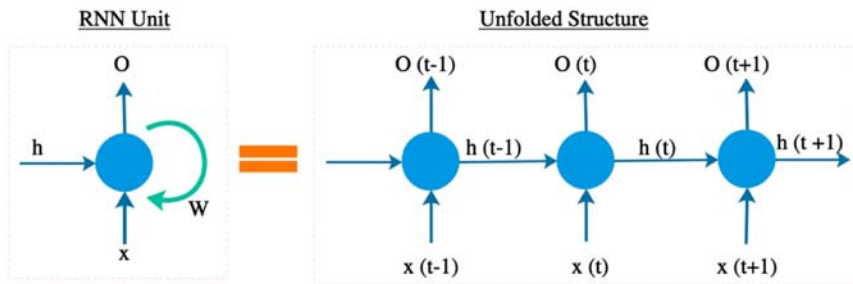


FIGURE 34.12 RNN unit (on the left) and unfolded structure (on the right). RNN, Recurrent Neural Networks. Adapted from Ndikumana, E., Minh, D.H.T., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. Remote Sensing, 10, 1217. doi:10.3390/rs10081217.

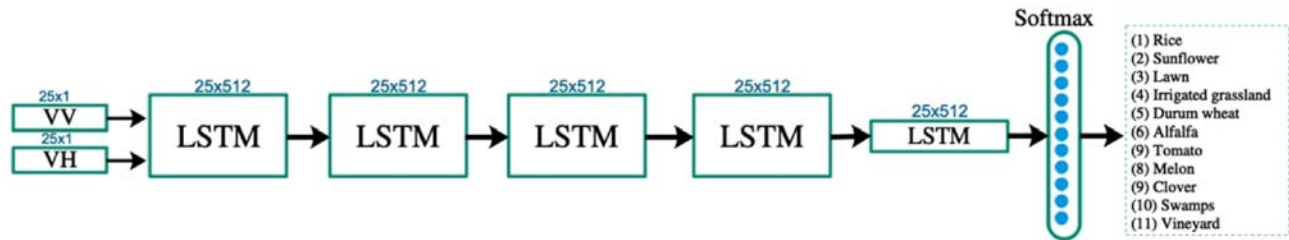


FIGURE 34.13 The schematic view of the RNN LSTM-based architecture. By replacing LSTM to GRU unit, we get the RNN GRU-based architecture. LSTM, long short-term memory; RNN, Recurrent Neural Networks. Adapted from Ndikumana, E., Minh, D.H.T., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. Remote Sensing, 10, 1217. doi:10.3390/rs10081217.

TABLE 34.2 The average and standard deviation from cross-validation five times on the time series SAR Sentinel-1 data.

Classifier	F-measure	Accuracy	Kappa
KNN	86.1 _ 0.6%	85.6 _ 0.6%	0.823 _ 0.009
Random forest	87.1 _ 0.9%	86.9 _ 1.2%	0.833 _ 0.015
Support vector machine	87.3 _ 1.5%	87.1 _ 1.6%	0.837 _ 0.019
RNN (LSTM)	89.2 _ 1.7%	89.1 _ 1.6%	0.862 _ 0.020
RNN (GRU)	89.8 _ 1.6%	89.6 _ 1.6%	0.869 _ 0.019

The higher values are in bold. KNN, K-nearest neighbor; LSTM, long short-term memory; RNN, Recurrent Neural Networks. Based on Ndikumana, E., Minh, D.H.T., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. Remote Sensing, 10, 1217. doi:10.3390/rs10081217.

The Sentinel-1 multitemporal data, after processed, was used as input for classification using classical approaches (KNN, RF, and SVM) and two models based on RNN (LSTM and GRU). The results of these different classification approaches are summarized in Table 34.2. This result is the performance of the fivefold cross-validation in the data from the Sentinel SAR-1 time series, showing the mean and standard deviation values of measure F (F-measure), accuracy, and Kappa assessment metrics from five repetitions. All classifier performance metrics for the multitemporal SAR Sentinel-1 data were very high, showing the quality of the dataset for agricultural classification tasks (Ndikumana et al., 2018).

Among the two RNN models, the GRU method obtained a slightly better result than the LSTM. This result is expected because the GRU unit is considered an improvement on the LSTM unit. Finally, applying the best classifier (RNN-based GRU) to the entire study of the area, the agricultural land cover map for Camargue was established in 2017 (Fig. 34.14). Fig. 34.15 is an enlarged version of the box with a white border in Fig. 34.14 to facilitate the visualization of the classification results for the RNN-based GRU and the SVM approach with the reference plots.

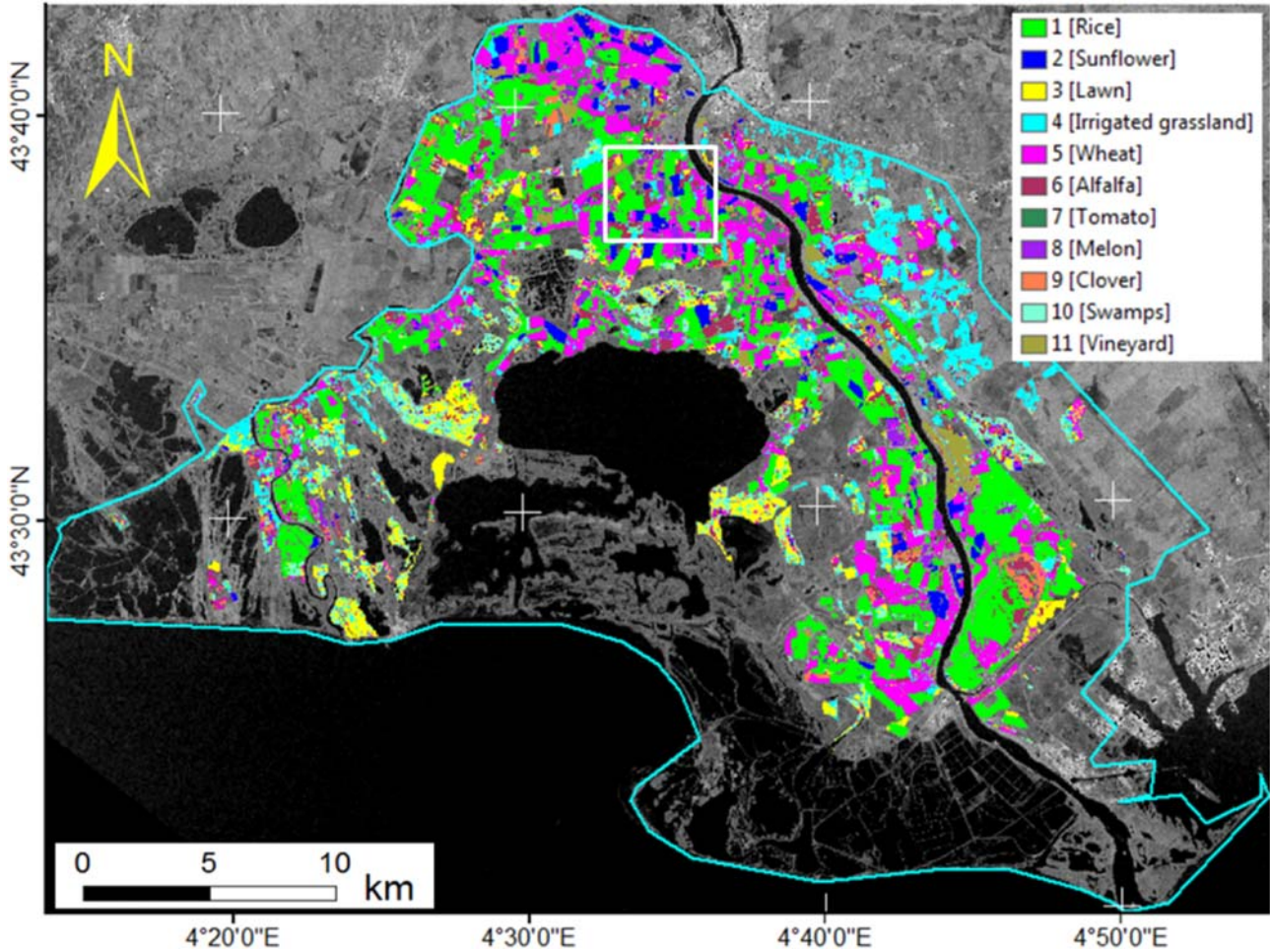


FIGURE 34.14 The agricultural land cover map in Camargue using the RNN-based GRU multitemporal SAR Sentinel-1. *RNN*, Recurrent Neural Networks. Based on Ndikumana, E., Minh, D.H.T., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, 10, 1217. doi:10.3390/rs10081217.

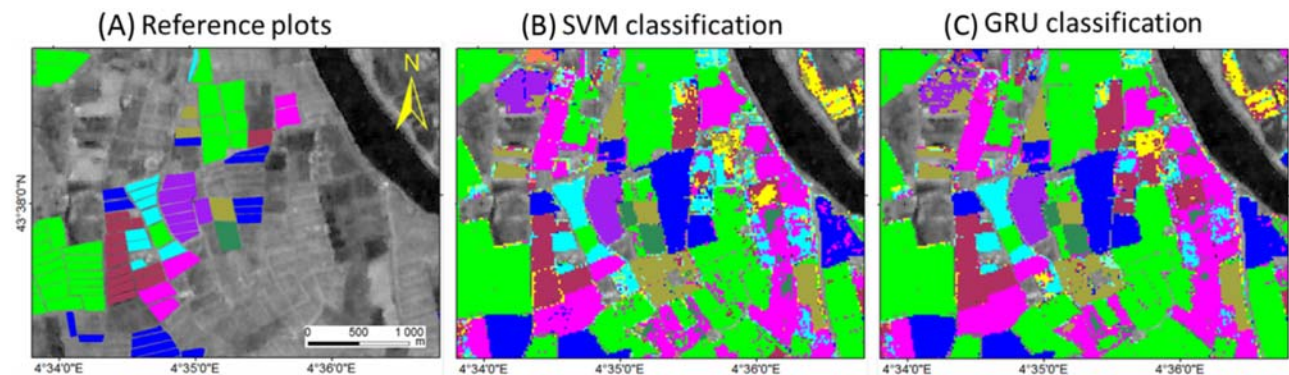


FIGURE 34.15 A zoom version of the white-border box in Fig. 34.14 is provided to facilitate visualization of classification results. (A) Reference plots; (B) the classical SVM result, and (C) the RNN-based GRU result. *RNN*, Recurrent Neural Networks; *SVM*, support vector machine. Based on Ndikumana, E., Minh, D.H.T., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, 10, 1217. doi:10.3390/rs10081217.

34.5 Conclusion

In recent years, it has been possible to observe a development in the computational techniques used to implement intelligent mechanisms in several application areas. Likewise, technological growth in areas such as Biological Engineering, as well as the evolution of sequencing platforms, was notorious. This is expected, since a sequencer is not created where there are no algorithms capable of processing the data produced.

Intelligent techniques known as Deep Learning, a more precise and improved form of Machine Learning, have now become an essential approach to a wide range of real-world problems, such as object perception, speech recognition, computer vision, collaborative filtering, and natural language processing. As more data are available, the system is able to learn the problem and provide solutions such as analysis and prediction of situations and behaviors.

DNNs are already present and are of great importance on bioinformatics and computational biology research due to their ability to handle complex and high-dimensional data. In its various architectures, such as Deep Autoencoder, RNN, DBN, Deep Boltzmann Machine, and CNN, these models are able to find correlations between previously unknown data. Today, there are many solutions and tools to apply Deep Learning to computational biology problems, such as predicting protein structures, regulating gene expression, and predicting diseases. They are also used in drug discovery, gene annotation, medical image recognition, and health-care management. Research continues to be carried out in this field to improve the efficiency of Deep Learning architectures. Bioinformatics and computational biology tend to continue to improve with the use of these techniques.

There is still a lot of space to expand and apply Deep Learning in agricultural research. Although some of the results have achieved accuracy at or above 95%, robustness and reliability are still challenges. The promise of applying Deep Learning in agriculture can be predicted. Furthermore, it is very likely that the future development of Deep Learning in the agricultural sciences will be based on the combination of various methods and techniques.

What is expected, in the future, with the evolution of computing and molecular biology, is that both can increasingly help the search for solutions to problems related to human, environmental, and animal health. There will be even greater impact with the integration of other sciences, such as chemistry and physics, in favor of knowledge and preservation of the environment, through monitoring by using One Health approaches.

References

- Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S., & Kruglyak, L. (2014). Genetics of single-cell protein abundance variation in large yeast populations. *Nature*, *506*, 494–497.
- Alcantara, G. (2017). Empirical analysis of non-linear activation functions for Deep Neural Networks in classification tasks. CoRR arXiv 2017, arXiv:1710.11272.
- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. Available from <https://doi.org/10.1038/nbt.3300>.
- Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., Zhavoronkov, A., & Albuquerque, N. (2016). Deep learning applications for predicting pharmacological properties of drugs. *Molecular Pharmaceutics*, *13*(7), 2524–2530. Available from <https://doi.org/10.1021/acs.molpharmaceut.6b00248>.Deep.
- Almeida, O.C.P. (2014). *Classificação de Tábuas de Madeira Usando Processamento de Imagens Digitais e Aprendizado de Máquina*. 2014. 105 f. Tese (Doutorado em Agronomia - Energia na Agricultura) -Faculdade de Ciências Agrônômicas, Universidade Estadual Paulista, Botucatu. Disponível em: <<http://hdl.handle.net/11449/115579>>. Acesso em: 27 nov. 2018.
- Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., & Powell, J. E. (2019). ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, *20*(1), 1–17. Available from <https://doi.org/10.1186/s13059-019-1862-5>.
- Amara, J., Bouaziz, B., & Algergawy, A. (2017). A Deep Learning-based Approach for Banana Leaf Diseases Classification. *BTW*.
- Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*, 878. Available from <https://doi.org/10.15252/msb.20156651>.
- Asgari, E., & Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*, *10*(11), 1–15. Available from <https://doi.org/10.1371/journal.pone.0141287>.
- Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics (Oxford, England)*, *15*(11), 937–946. Available from <https://doi.org/10.1093/bioinformatics/15.11.937>.
- Bastiaanssen, W. G. M., Molden, D. J., & Makin, I. W. (2000). Remote sensing for irrigated agriculture: Examples from research and possible applications. *Agricultural Water Management*, *46*, 137–155.
- Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., & Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. *Science (New York, N.Y.)*, *347*, 664–667.
- Bargoti, S., & Underwood, J. (2017). Deep fruit detection in orchards. In A. Okamura (Ed.), *In 2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3626–3633). Piscataway, NJ, USA: IEEE.

- Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., . . . Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, 12, R10.
- Bengio, Y., Courville, A. C., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE TPAMI*, 35, 1798–1828.
- Bragamonte, J. S., Camargo, S. S., Cardoso, L. L., Yokoo, M. J. I., & Cardoso, F. F. (2018). *Estimación Automática de Espesura de Gordura Subcutánea Bovina em Imagens Ultrassonográficas Utilizando Deep Learning*. 47JAIIO – CAI (Congreso Argentino de AgroInformática) - ISSN: 2525-0949.
- Brosnan, T., & Sun, D. (2002). Inspection and grading of agricultural and food products by computer vision systems – A review. *Computers and Electronics in Agriculture*, 193–213.
- Buckley, C., & Carney, P. (2013). The potential to reduce the risk of diffuse pollution from agriculture while improving economic performance at farm level. *Environmental Science and Policy*, 25, 118–126.
- Buduma, N., & Locascio, N. (2017). *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms* (1st edn.). O'Reilly Media, Inc, [S.l.].
- Cardoso, L. L. (2013). *Estimativas do Rendimento Comercial de Novilhos com a Utilização de Ultrassom*. Tese (doutorado) - Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia, Programa de Pós-Graduação em Zootecnia, Porto Alegre.
- Carvalho, F. P. (2006). Agriculture, pesticides, food security and food safety. *Environmental Science and Policy*, 9, 685–692.
- Cerutti, S., Baselli, G., Bianchi, A. M., Caiani, E., Contini, D., Cubeddu, R., . . . Torricelli, A. (2011). Biomedical signal and image processing. *IEEE Pulse*, 2(3), 41–54. Available from <https://doi.org/10.1109/MPUL.2011.941522>.
- Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6), 1248–1259. Available from <https://doi.org/10.1158/1078-0432.CCR-17-0853>.
- Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7, 2094–2107.
- Chen, S. W., Shivakumar, S. S., Dcunha, S., Das, J., Okon, E., Qu, C., . . . Kumar, V. (2017). Counting apples and oranges with deep learning: A data-driven approach. *IEEE Robotics and Automation Letters*, 2, 781–788.
- Chen, Y., Li, Y., Narayan, R., Subramanian, A., & Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics (Oxford, England)*, 32(12), 1832–1839. Available from <https://doi.org/10.1093/bioinformatics/btw074>.
- Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J., & Zhu, Y. (2016). Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 104, 2207–2219.
- Cho, K., van Merriënboer, B., Gülehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the EMNLP*, Doha, Qatar, 25–29 October 2014, pp. 1724–1734.
- Chollet, F. (2017). *Deep learning with python* (1st edn.). Greenwich, CT: Manning Publications Co.
- Christiansen, P., Nielsen, L. N., Steen, K. A., Jørgensen, R. N., & Karstoft, H. (2016). Deep anomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field. *Sensors*, 16, E1904.
- Ciresan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, pp. 411–418. https://doi.org/10.1007/978-3-642-40763-5_51.
- de Almeida, M. H. B., Gomes, R. C., de Almeida, O. C. P., & Ballarin, A. W. (2018). Desempenho da Técnica Deep Learning na Análise e Categorização de Imagens de Defeito de Madeira. *Energia na Agricultura*, 33(3), 284–291.
- Dediu, A. H., Hernández-Quiroz, F., Martín-Vide, C., & Rosenblueth, D. A. (2015). Convolutional LSTM Networks for Subcellular Localization of Proteins. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9199. <https://doi.org/10.1007/978-3-319-21233-3>.
- Denas, O., & Taylor, J. (2013). Deep modeling of gene expression regulation in an Erythropoiesis model. *ICML workshop on representation learning*.
- D'haeseleer, P. (2006). What are DNA sequence motifs? *Nature Biotechnology*, 24, 423–425. Available from <https://doi.org/10.1038/nbt0406-423>.
- Douarre, C., Schielein, R., Frindel, C., Gerth, S., and Rousseau, D. (2016). Deep learning based root-soil segmentation from X-ray tomography. bioRxiv, 071662. <https://doi.org/10.1101/071662>.
- Drusch, M., Bello, U. D., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al. (2012). Sentinel-2 ESA optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120, 25–36.
- Dyrmann, M., Karstoft, H., & Midtby, H. S. (2016). *Plant species classification using deep convolutional neural network*. *Biosystems Engineering*, . *FAO (2009) How to Feed the World in 2050* (151, pp. 72–80). Rome, Italy: FAO.
- Eduati, F., Mangravite, L. M., Wang, T., Tang, H., Bare, J. C., Huang, R., Norman, T., Kellen, M., Menden, M. P., Yang, J., Zhan, X., Zhong, R., Xiao, G., Xia, M., Abdo, N., Kosyk, O., Collaboration, N.-N.-U. D. T., Friend, S., Dearry, A., Simeonov, A., et al. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. *Nature Biotechnology*, 33, 933–940.
- Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1), 1–14. Available from <https://doi.org/10.1038/s41467-018-07931-2>.
- FAO. (2009). *How to feed the world in 2050*. Rome, Italy: FAO.
- Ferreira, A. S. (2017). *Redes Neurais Convolucionais Profundas na Detecção de Plantas Daninhas em Lavoura de Soja*. Dissertação de Mestrado. Campo Grande: UFMS.
- Flamary, R., Fauvel, M., Mura, M. D., & Valero, S. (2015). Analysis of multitemporal classification techniques for forecasting image time series. *IEEE Geoscience and Remote Sensing Letters*, 12, 953–957.

- Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., et al. (2005). Global consequences of land use. *Science (New York, N.Y.)*, 309, 570–574.
- Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61, 399–409.
- Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., et al. (2018). From hype to reality: Data science enabling personalized medicine. *BMC Medicine*, 16, 1–15. Available from <https://doi.org/10.1186/s12916-018-1122-7>.
- Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science (New York, N.Y.)*, 327, 828–831.
- Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., . . . Troncoso, J. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genetics*, 6, e1000952.
- Gomez-Chova, L., Camps-Valls, G., Munoz-Mari, J., & Calpe, J. (2008). Semisupervised image classification with laplacian support vector machines. *IEEE Geoscience and Remote Sensing Letters*, 5, 336–340.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press, [S.l.].
- Grapov, D., Fahrman, J., Wanichthanarak, K., & Khoomrung, S. (2018). Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *OMICS a Journal of Integrative Biology*, 22(10), 630–636. Available from <https://doi.org/10.1089/omi.2018.0097>.
- Graves, A., Mohamed, A., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 26–31 May 2013, pp. 6645–6649.
- Green, (E. D.), Watson, J. D., & Collins, F. S. (2015). Twenty-five years of big biology. *Nature*, 526, 29–31. Available from <https://doi.org/10.1007/s13398-014-0173-7.2>.
- Greiner, S. (2012) Chapter III – Ultrasound and the beef carcass. In *Ultrasound guidelines council field technician study guide, 2012 edition*.
- Grinblat, G. L., Uzal, L. C., Larese, M. G., & Granitto, P. M. (2016). Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture*, 127, 418–424.
- Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., . . . Snyder, M. (2015). Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162, 1051–1065.
- Gu, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354–377.
- Guan, R., Wang, X., Yang, M. Q., Zhang, Y., Zhou, F., Yang, C., & Liang, Y. (2018). Multi-label deep learning for gene function annotation in cancer pathways. *Scientific Reports*, 8(1), 1–9. Available from <https://doi.org/10.1038/s41598-017-17842-9>.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. (2015). The rise of ‘big data’ on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv:1512.03385.
- Heffernan, R., Paliwal, K., Lyons, J., Dehngi, A., Sharma, A., Wang, J., . . . Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, 5(March), 1–11. Available from <https://doi.org/10.1038/srep11476>.
- Ho Tong Minh, D., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F., & Maurel, P. (2018). Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1. *IEEE Geoscience and Remote Sensing Letters*, 15, 464–468.
- Hochreiter, S., Heusel, M., & Obermayer, K. (2007). Fast model-based protein homology detection without alignment. *Bioinformatics (Oxford, England)*, 23(14), 1728–1736. Available from <https://doi.org/10.1093/bioinformatics/btm247>.
- Hochreiter, S., & Schmidhuber, J. (1996). LSTM can solve hard long time lag problems. In *Proceedings of the NIPS*, Denver, CO, 2–5 December 1996, pp. 473–479.
- Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*, 50(8). Available from <https://doi.org/10.1038/s12276-018-0071-8>.
- Ienco, D., Gaetano, R., Dupaquier, C., & Maurel, P. (2017). Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14, 1685–1689.
- Inglada, J., Vincent, A., Arias, M., & Marais-Sicre, C. (2016). Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. *Remote Sensing*, 8, 362.
- Ishimwe, R., Abutaleb, K., & Ahmed, F. (2014). Applications of thermal imaging in agriculture – A review. *Advances in Remote Sensing*, 3, 128–140.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., . . . Darrel, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on multimedia*, pp. 675–678. <https://doi.org/10.1145/2647868.2654889>.
- Kamilaris, A., Gao, F., Prenafeta-Boldú, F. X., & Ali, M. I. (2016). Agri-IoT: A semantic framework for Internet of Things-enabled smart farming applications. In *3rd World Forum on Internet of Things (WF-IoT)*, Reston, VA, IEEE, pp. 442–447.
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2017). Disaster monitoring using unmanned aerial vehicles and deep learning. In *Disaster Management for Resilience and Public Safety Workshop, Proceedings of EnviroInfo 2017*, Luxembourg.
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2018a). A review of the use of convolutional neural networks in agriculture. *The Journal of Agricultural Science*, 156, 312–322. Available from <https://doi.org/10.1017/S0021859618000436>.
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2018b). *Deep learning in agriculture: A survey*. Technical Report. Institute for Food and Agricultural Research and Technology (IRTA).
- Kell, D. B. (2006). Metabolomics, modelling and machine learning in systems biology - Towards an understanding of the languages of cells. *FEBS Journal*, 273(5), 873–894. Available from <https://doi.org/10.1111/j.1742-4658.2006.05136.x>.

- Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7), 990–999. Available from <https://doi.org/10.1101/gr.200535.115>.
- Kitzes, J., Wackernagel, M., Loh, J., Peller, A., Goldfinger, S., Cheng, D., & Tea, K. (2008). Shrink and share: Humanity's present and future ecological footprint. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1491), 467–475.
- Kline, D. E., Surak, C., & Araman, P. A. (2018). Automated hardwood lumber grading utilizing a multiple sensor machine vision technology. *Computers and Electronics in Agriculture, Virginia*, 41(1/3), 139–155. <<http://www.sciencedirect.com/science/article/pii/S0168169903000486>>. Accessed 28.11.03.
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). Review the technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4), 610–620. Available from <https://doi.org/10.1016/j.molcel.2015.04.005>.
- Kuwata, K., & Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, Milan, Italy, pp. 858–861.
- Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14, 778–782.
- Le Toan, T., Ribbes, F., Wang, L. F., Floury, N., Ding, K. H., Kong, J. A., ... Kurosu, T. (1997). Rice crop mapping and monitoring using ERS-1 data based on experiment and modeling results. *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society*, 35, 41–56.
- Lee, S. H., Chan, C. S., Wilkin, P., & Remagnino, P. (2015). *Deep-plant: Plant identification with convolutional neural networks*. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 452–456). Piscataway, NJ, USA: IEEE.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *International Journal of Natural and Social Sciences*, 521, 436–444.
- Lee, T., & Yoon, S. (2015). Boosted categorical restricted Boltzmann machine for computational prediction of splice junctions. *32nd International Conference on Machine Learning, ICML 2015*, 3, 2473–2482.
- Leung, M. K. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics (Oxford, England)*, 30(12), 121–129. Available from <https://doi.org/10.1093/bioinformatics/btu277>.
- Li, C., Wang, J., Wang, L., Hu, L., & Gong, P. (2014). Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. *Remote Sensing*, 6, 964–983.
- Li, J., Bioucas-Dias, J. M., & Plaza, A. (2010). Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society*, 48, 4085–4098.
- Li, Y., Chen, C. Y., & Wasserman, W. W. (2016). Deep feature selection: Theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23(5), 322–336. Available from <https://doi.org/10.1089/cmb.2015.0189>.
- Li, Y., & Ngom, A. (2015). Data integration in machine learning. *Proceedings – 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015*, 1665–1671. <https://doi.org/10.1109/BIBM.2015.7359925>.
- Li, Y., Wu, F. X., & Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2), 325–340. Available from <https://doi.org/10.1093/bib/bbw113>.
- Liaghat, S., & Balasundram, S. K. (2010). A review: The role of remote sensing in precision agriculture. *American Journal of Agricultural and Biological Sciences*, 5, 50–55.
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18, 2674. Available from <http://doi.org/10.3390/s18082674>. Available from <http://www.mdpi.com/journal/sensors>.
- Liang, Z., Huang, J. X., Zeng, X., & Zhang, G. (2016). DL-ADR: A novel deep learning model for classifying genomic variants into adverse drug reactions. *BMC Medical Genomics*, 9(Suppl. 2). Available from <https://doi.org/10.1186/s12920-016-0207-4>.
- Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *TACL*, 4, 521–535.
- Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 1053–1058. Available from <https://doi.org/10.1038/s41592-018-0229-2>.
- Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2019). scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8), 715–721. Available from <https://doi.org/10.1038/s41592-019-0494-8>.
- Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28, 823–870.
- Lu, H., Fu, X., Liu, C., Li, L. G., He, Y. X., & Li, N. W. (2017). Cultivated land information extraction in UAV imagery based on deep convolutional neural network and transfer learning. *Journal of Mountain Science*, 14, 731–741.
- Luus, F. P., Salmon, B. P., van den Bergh, F., & Maharaj, B. T. (2015). Multiview deep learning for land-use classification. *IEEE Geoscience and Remote Sensing Letters*, 12, 2448–2452.
- Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., ... Yang, Y. (2014). Predicting backbone C α angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry*, 35(28), 2040–2046. Available from <https://doi.org/10.1002/jcc.23718>.
- Mandic, D. P., & Chambers, J. A. (2001). *Recurrent neural networks for prediction: Learning algorithms, architectures and stability*. New York: John Wiley.
- Manyika, J., Michael, C., Brad, B., Jacques, B., Richard, D., Charles, R., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute* (June), 156. Available from <https://doi.org/10.1080/01443610903114527>.

- Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One*, *8*, e61318.
- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, *7*, 1419.
- Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., ... Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, *464*, 773–777.
- Mouret, J. C. (1988). Etude de l'Agrosystème Rizicole en Camargue dans ses Relations avec le Milieu et le Systeme Cultural: Aspects Particuliers de la Fertilité. Ph.D. Thesis, Université des Sciences et Techniques du Languedoc, Montpellier, France.
- Munoz-Mari, J., Bovolo, F., Gomez-Chova, L., Bruzzone, L., & Camp-Valls, G. (2010). Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society*, *48*, 3188–3197.
- Ndikumana, E., Minh, D. H. T., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, *10*, 1217. Available from <https://doi.org/10.3390/rs10081217>.
- Oliveira, W. (2018). *Software para Reconhecimento de Espécies Florestais a Partir de Imagens Digitais de Madeiras Utilizando Deep Learning*. Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Tecnologias Computacionais para o Agronegócio. Medianeira.
- Ouzounis, C. A. (2012). Rise and demise of bioinformatics? promise and progress. *PLoS Computational Biology*, *8*. Available from <https://doi.org/10.1371/journal.pcbi.1002487>.
- Ozdogan, M., Yang, Y., Allez, G., & Cervantes, C. (2010). Remote sensing of irrigated agriculture: Opportunities and challenges. *Remote Sensing*, *2*, 2274–2304.
- Park, S., Min, S., Choi, H., & Yoon, S. (2016). deepMiRGene: Deep neural network based precursor microRNA prediction. *ArXiv Preprint ArXiv:1605.00017*.
- Park, Y., & Kellis, M. (2015). Deep learning for regulatory genomics. *Nature Biotechnology*, *33*(8), 825–826. Available from <https://doi.org/10.1038/nbt.3313>.
- Parts, L., Liu, Y. C., Tekkedil, M. M., Steinmetz, L. M., Caudy, A. A., Fraser, A. G., ... Rosebrock, A. P. (2014). Heritability and genetic basis of protein level variation in an outbred population. *Genome Research*, *24*, 1363–1370.
- Pearson, W. R. (2001). Training for bioinformatics and computational biology. *Bioinformatics (Oxford, England)*, *17*, 761–762. Available from <https://doi.org/10.1093/bioinformatics/17.9.76>.
- Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*, 768–772.
- Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., & Wishart, D. (2019). Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites*, *9*(4), 1–31. Available from <https://doi.org/10.3390/metabo9040076>.
- Plis, S. M., Hjelm, D. R., Slakhtudinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., ... Calhoun, V. D. (2014). Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience*, *8*(8 July), 1–11. Available from <https://doi.org/10.3389/fnins.2014.00229>.
- Polso, A., Speedy, A., & Kueneman, E. (2004). Good agricultural practices—A working concept. In *Proceedings of the FAO Internal Workshop on Good Agricultural Practices*, Rome, Italy, 27–29 October 2004; Vol. 1, p. 41.
- Prats-Iraola, P., Scheiber, R., Marotti, L., Wollstadt, S., & Reigber, A. (2012). TOPS interferometry with TerraSAR-X. *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society*, *50*, 3179–3188.
- Rahmoonfar, M., & Sheppard, C. (2017). Deep count: Fruit counting based on deep simulated learning. *Sensors*, *17*, 905.
- Rall, R. (2010). *Processamento de Imagens Digitais para Detecção e Quantificação de Defeitos na Madeira Serrada de Coníferas de Reflorestamento de Uso não Estrutural*. 2010. 123 f. Tese (Doutorado em Agronomia - Energia na Agricultura) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu. Disponível em: <<http://hdl.handle.net/11449/101882>>. Accessed 27.11.18.
- Relling, M. V., & Evans, W. E. (2015). *Pharmacogenomics in the clinic*. *Nature*. Nature Publishing Group. Available from <https://doi.org/10.1038/nature15817>.
- Reyes, A. K., Caicedo, J. C., & Camargo, J. E. (2015). Fine-tuning deep convolutional networks for plant recognition. In L. Cappellato, N. Ferro, G. J. F. Jones, & E. San Juan (Eds.), *CLEF2015 Working Notes. Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum*, Toulouse, France, 8–11 September 2015. Toulouse: CLEF. Available online from: <http://ceur-ws.org/Vol-1391/>. Accessed 11.06.18.
- Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). *Methods of integrating data to uncover genotype-phenotype interactions*. *Nature Reviews Genetics*. Nature Publishing Group. Available from <https://doi.org/10.1038/nrg3868>.
- Robertson, G., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, *4*(8), 651–657.
- Ronen, J., Hayat, S., & Akalin, A. (2019). Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Science Alliance*, *2*(6), 1–16. Available from <https://doi.org/10.26508/lsa.201900517>.
- Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., ... Summers, R. M. (2016). Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Transactions on Medical Imaging*, *35*(5), 1170–1181. Available from <https://doi.org/10.1109/TMI.2015.2482920>.
- Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., & McCool, C. (2016). Deepfruits: A fruit detection system using deep neural networks. *Sensors*, *16*, E1222.

- Saha, S., Sengupta, K., Chatterjee, P., Basu, S., & Nasipuri, M. (2018). Analysis of protein targets in pathogen-host interaction in infectious diseases: A case study on *Plasmodium falciparum* and *Homo sapiens* interaction network. *Briefings in Functional Genomics*, *17*, 441–450. Available from <https://doi.org/10.1093/bfpg/elx024>.
- Santoni, M. M., Sensuse, D. I., Arymurthy, A. M., & Fanany, M. I. (2015). Cattle race classification using gray level co-occurrence matrix convolutional neural networks. *Procedia Computer Science*, *59*, 493–502.
- Saxena, L., & Armstrong, L. (2014). A survey of image processing techniques for agriculture. In *Proceedings of Asian Federation for Information Technology in Agriculture*. Australian Society of Information and Communication Technologies in Agriculture, Perth, Australia, pp. 401–413.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.
- Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*, 16–18. Available from <https://doi.org/10.1038/NMETH1156>.
- Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience*, *2016*, 3289801, 11 p. Hindawi Publishing Corporation. <https://doi.org/10.1155/2016/3289801>.
- Soma, K., Mori, R., Sato, R., Furumai, N., & Nara, S. (2015). Simultaneous multichannel signal transfers via chaos in a recurrent neural network. *Neural Computation*, *27*, 1083–1101.
- Song, X., Zhang, G., Liu, F., Li, D., Zhao, Y., & Yang, J. (2016). Modeling spatiotemporal distribution of soil moisture by deep learning-based cellular automata model. *Journal of Arid Land*, *8*, 734–748.
- Sørensen, R. A., Rasmussen, J., Nielsen, J., & Jørgensen, R. N. (2017). *Thistle Detection Using Convolutional Neural Networks*. Montpellier, France: EFITA Congress.
- Spencer, M., Eickholt, J., & Cheng, J. (2015). A deep learning network approach to. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *12*(1), 103–112. Available from <https://doi.org/10.1109/TCBB.2014.2343960>.
- Steen, K. A., Christiansen, P., Karstoft, H., & Jørgensen, R. N. (2016). Using deep learning to challenge safety standard for highly autonomous machines in agriculture. *Journal of Imaging*, *2*, 6.
- Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., . . . Teichmann, S. A. (2017). Power analysis of single-cell mRNA-sequencing experiments. *Nature Methods*, *14*(4), 381–387. Available from <https://doi.org/10.1038/nmeth.422>.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Piscataway, NJ, pp. 1–9.
- Teke, M., Deveci, H. S., Haliloğlu, O., Gürbüz, S. Z., & Sakarya, U. (2013). A short survey of hyperspectral remote sensing applications in agriculture. In M. Ilarslan, F. Ince, O. Kaynak, & S. Basturk (Eds.), *6th International Conference on Recent Advances in Space Technologies (RAST)*, IEEE (pp. 171–176). Piscataway, NJ: IEEE.
- Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., et al. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9–24.
- Tuia, D., Volpi, M., Copa, L., Kanevski, M., & Munoz-Mari, J. (2011). A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal on Selected Topics in Signal Processing*, *5*, 606–617.
- Tyagi, A. C. (2016). Towards a second green revolution. *Irrigation and Drainage*, *65*, 388–389.
- Waga, D., & Rabah, K. (2014). Environmental conditions' big data management and cloud computing analytics for sustainable agriculture. *World Journal of Computer Application and Technology*, *2*, 73–81.
- Waske, B., & Braun, M. (2009). Classifier ensembles for land cover mapping using multitemporal SAR imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *64*, 450–457.
- Waszak, S. M., Delaneau, O., Gschwind, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., Orioli, A., Wiederkehr, M., Panousis, N. I., Yurovsky, A., Romano-Palumbo, L., Planchon, A., Bielser, D., Padiouleau, I., Udin, G., Thurnheer, S., Hacker, D., Hernandez, N., Reymond, A., Deplancke, B., et al. (2015). Population variation and genetic control of modular chromatin architecture in humans. *Cell*, *162*, 1039–1050.
- Way, G. P., & Greene, C. S. (2018). Bayesian deep learning for single-cell analysis. *Nature Methods*, *15*(12), 1009–1010. Available from <https://doi.org/10.1038/s41592-018-0230-9>.
- Weber, R. H., & Weber, R. (2010). *Internet of things* (Vol. 12). New York, NY: Springer.
- Woalder. (2017). Stacked Sparse Autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *Physiology & Behavior*, *176*(1), 139–148. Available from <https://doi.org/10.1016/j.physbeh.2017.03.040>.
- Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., . . . Frey, B. J. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science (New York, N.Y.)*, *347*, 1254806.
- Xinshao, W., & Cheng, C. (2015). *Weed seeds classification based on PCANet deep learning baseline*. In *IEEE Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 408–415). Hong Kong, China: Asia-Pacific Signal and Information Processing Association.
- Yokoo, M. J.-I., et al. (2009). Correlações Genéticas entre Escores Visuais e Características de Carcaça Medidas por Ultrassom em Bovinos de Corte. *Pesquisa Agropecuária Brasileira*, *44*(2), 197–202.
- Yokoo, M. J.-I., et al. (2015). *Avaliação de Carcaça por Ultrassom e sua Aplicação Prática: Qual é a Importância desta Tecnologia para o Produtor*. Bagé: Embrapa Pecuária Sul.
- Ypsilantis, P. P., Siddique, M., Sohn, H. M., Davies, A., Cook, G., Goh, V., & Montana, G. (2015). Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks. *PLoS One*, *10*(9), 1–18. Available from <https://doi.org/10.1371/journal.pone.0137036>.

- Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., . . . Shi, T. (2018). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Frontiers in Genetics*, 9(October), 1–9. Available from <https://doi.org/10.3389/fgene.2018.00477>.
- Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., & Peng, S. (2019). Deep learning in omics: A survey and guideline. *Briefings in Functional Genomics*, 18(1), 41–57. Available from <https://doi.org/10.1093/bfpg/ely030>.
- Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A., Jin, W., & Schuller, B. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology*, 9, 49.
- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12, 931–934.
- Zhou, Y., & Chellappa, R. (1988). Computation of optical flow using a neural network. In *Proceedings of the IEEE 1988 International Conference on Neural Networks*, San Diego, CA, pp. 24–27.
- Zhu, N. Y., et al. (2018). Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *International Journal of Agricultural and Biological Engineering*, 11(4), Open Access at. Available from <https://www.ijabe.org>.

Image processing–based artificial intelligence system for rapid detection of plant diseases

Sanjaya Shankar Tripathy¹, Raju Poddar², Lopamudra Satapathy³ and Kunal Mukhopadhyay²

¹Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Jharkhand, India, ²Department of Bioengineering and Biotechnology, Birla Institute of Technology, Mesra, Jharkhand, India, ³Faculty of Agriculture, Usha Martin University, Ranchi, Jharkhand, India

35.1 Introduction

Artificial intelligence (AI) is contributing significantly to all domains of the industry. Every sector is looking for automation of certain jobs using intelligent machinery. Agriculture plays a vital role in the economic sector. Worldwide, agriculture is a \$5 trillion industry. The global population is expected to reach more than 9 billion by 2050 which will require an increase in agricultural production by 70% to fulfill the demand. Due to increase in world population; land, water, and resources are becoming insufficient to maintain the demand–supply chain. Hence, we need a smarter approach in crop production and storage. The life cycle of agriculture comprises soil preparation, sowing seeds, application of fertilizer, irrigation, weed management, disease management, crop protection, harvesting and storage. AI and machine learning (ML) can be applied in all these steps of agriculture to increase the productivity. In this chapter, disease management using AI and ML is explored. In agriculture, disease detection plays an important role. Early detection of plant disease will help the farmer to protect the plant from the disease. Manual identification is tedious and time-consuming. It also requires expertise in the specific crop/plant. Timely availability of the experts becomes difficult in remote locations; hence automatic identification of plant disease will certainly be able to help the farmer in identification of disease at the earliest. Plant disease can be identified from different sections of plant such as root, stem, leaf, flower, and fruit. In plants, some common diseases are seen such as brown and yellow spots, early and late scorch, and other fungal, viral, and bacterial diseases. Image processing is a tool to identify the affected area of disease and determine the difference in the color of the affected area (Dhaygude & Kumbhar, 2013; Ghaiwat & Arora, 2014). Nowadays smartphones can offer novel approaches to identify diseases because of their high-resolution cameras, good network facility, computing power, long battery life, and high-resolution displays.

35.2 Visual symptoms of diseases in plant

To identify the plant disease, one should be able to identify the healthy plant. Growth rate, color, texture, and shapes of the leaves differ from plant to plant.

Prior knowledge is required to know the difference between the normal appearance of a plant and the appearance of different cultivars. Once the “normal” appearance of the plant under consideration is determined, comparisons can be made to identify the infected plant. Around 85% of the plant diseases are caused by the three main pathogenic microbes: virus, bacteria, and fungus.

Microbial pathogens like fungus, bacteria, and virus are mostly responsible for causing diseases in plant. Symptoms of plant disease are a physical evidence of the pathogen. For example, fungal fruiting bodies are a sign of disease. The appearance of powdery mildew on a lilac leaf, actually, shows the parasitic fungal disease organism itself

(*Microsphaera alni*). A symptom of plant disease is a visible effect of the disease on the plant. Generally, plants respond to the pathogens by changing its color, texture, and shape of the leaf. This visible change gives a clue about the disease in the plant.

Symptoms may include a detectable change in color, shape, or function of the plant as it responds to the pathogen. Symptoms can be grouped as follows:

- unnatural growth of tissues and organs
- undeveloped tissues and organs
- dead of leaves/stem
- change in the color/texture of the leaves/stem

The infected area of the crop can appear at leaves, stems, roots, flowers, and fruits. Flower and fruits come to the plant in a later stage of the plant development. Visual examination of the root is not possible since roots are deep into the earth surface. In a plant, we have multiple leaves but the plants have a very limited number of stems. So for early detection of disease, leaf is considered the best choice. More than 50% of the disease symptom comes on leaves only. The health of leaf defines the degree of healthiness of the plant.

35.3 Imaging

During the various stresses, including pathogen attack, plants counter it with biochemical and biophysical changes—different phenomena like changes in microstructure of leaf and degradation of chlorophyll content are found. Several existing imaging methods are available like hyperspectral reflection, multispectral fluorescence, and optical coherence tomography (OCT) for imaging different plant parts under field conditions.

Hyperspectral imaging relies on measuring and analyzing of reflected light pattern in narrow bands spectrum as a hypercube (Moghadam et al., 2017). It can be used as detection as plant disease characterization and classification. However, different parts of plants interacts with different bands of electromagnetic spectrum based on biochemical compounds and microstructure. An example, leaf of healthy plants absorbs visible range of light (400–700 nm) due to the presence of photosynthesis pigments. However, the spectrum ranges from visible to shortwave infrared (400–2500 nm) based on water and chemical contents inside of a leaf (Fig. 35.1B).

On the other side, imaging system OCT generates tomographic images of the plant leaf (Rateria, Mohan, Mukhopadhyay, & Poddar, 2019; Wijesinghe et al., 2016). The OCT B-scan cross-sectional images reveal the changes in internal structure, in real-time, in vivo, fast, and noninvasive ways (Fig. 35.1C and D).

A huge pool of such images can be generated using these abovementioned imaging systems for infected and normal plants.

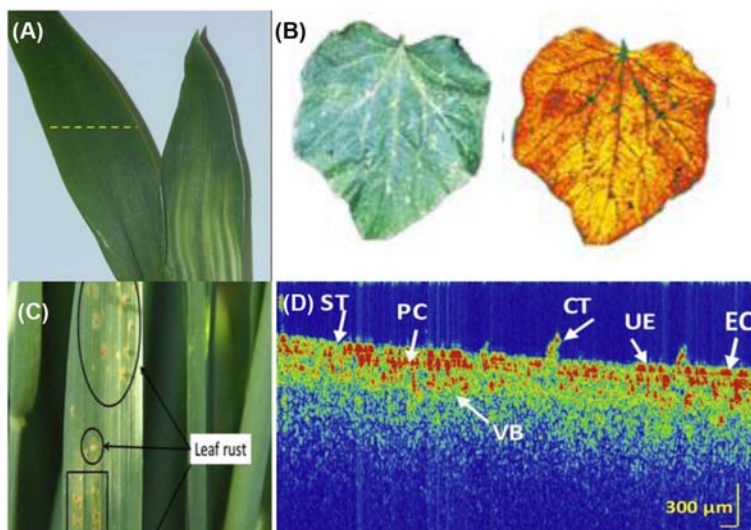


FIGURE 35.1 (A) Uninfected leaves; (B) hyperspectral image of infected leaves; (C) infected wheat leaves; and (D) real-time, in-vivo OCT B-scan. UE, upper epidermis; PC, parenchyma cell; VB, vascular bundle; EC, epidermal cell; ST, stomata. Yellow dashed line (in A) shows the position of OCT scan (D) [Rateria et al (2019)].

35.4 Database creation

The success of the plant disease identification depends heavily on the datasets. Image datasets are created by capturing images of the leaf of different crops under consideration. This database should contain healthy images and images of the leaf infected by different diseases. These images are also taken in different lighting conditions.

35.5 Disease identification using feature extraction and classification

Steps involved in disease identification using feature extraction are shown in Fig. 35.2.

First, the infected leaf image is acquired by using any digital camera. Preprocessing steps such as background removal and noise reduction are done on the acquired image. Then, segmentation is done to get only the infected portion of the leaf. Color and texture features were extracted from the segmented image. This process is done for all the images present in the database. The image dataset is split into training set and validation/testing set. With the help of training dataset, disease identification system is trained using different ML algorithms such as support vector machine (SVM), K-nearest neighbour (KNN), K-means, and random forest. The learning algorithm can be supervised learning, unsupervised learning, and reinforcement learning. Once the training is satisfactory, the system is tested using validation dataset.

Identification and classification of grapevine diseases is proposed by Meunkaewjinda, Kumsawat, Attakitmongcol, and Srikaew (2008). Different color spaces such as HIS, YCbCr, $L^*a^*b^*$, and UVL were explored in their execution. The background removal was performed by a multilayer perceptron network, which is coupled with a color library built a priori by means of an unsupervised self-organizing map (SOM). The color patches on the leaves were then clustered by unsupervised and untrained SOM. Genetic algorithm was used to identify the number of clusters to be adopted in each case. Healthy and diseased regions were then separated by an SVM. After applying a specific threshold the segmented image was submitted to a multiclass SVM, which performed the classification into either healthy, scab, or rust diseases.

Cucumber disease recognition system based on image processing and SVM was proposed (Youwen, Tianlai, & Yan, 2008). They developed a method to identify two diseases that are predominantly present in cucumber leaves. Statistical pattern recognition technique was adapted for segmentation of healthy and diseased region. Color, texture, and shape

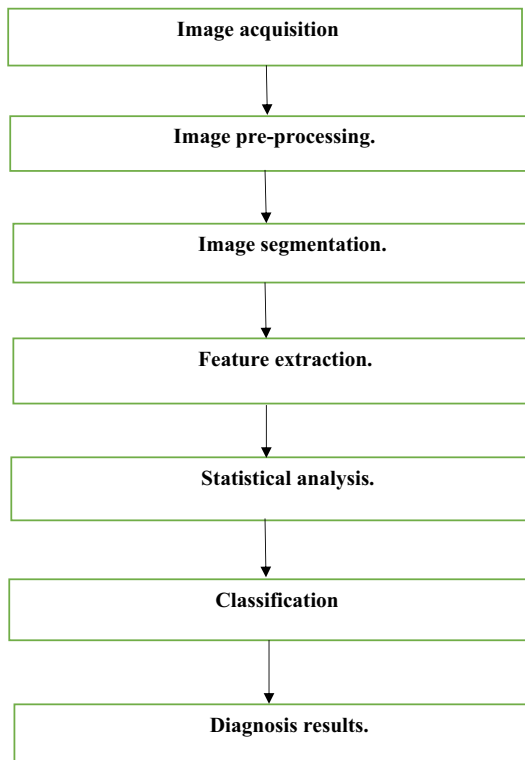


FIGURE 35.2 Different steps of disease detection.

features were extracted and feed into an SVM, which performed the disease classification. The authors stated that the results obtained by them using SVM were much better than that of ANN. Yao et al. (2009) applied SVM for disease identification and classification in rice crops. First, the RGB image was converted into HSV color space using color transformation. Diseased regions were extracted from the whole image using segmentation by Otsu's method. Color, texture, and shape features were extracted from the HSV-transformed image. These features were applied to SVM, for the final disease classification. Identification and classification of disease-causing agents in cotton plants was developed (Camargo & Smith, 2009). The system could able to identify three different diseases in cotton plant. In their study, they incorporated fruits and stem along with the leaves to identify the disease very accurately. They also applied SVM for classification of disease. Inputs to the SVM are the extracted features from the infected region and the output is the disease class. Hsu and Lin (2002) used SVM to deal with multiclass disease. The authors concluded that the texture features have the best discriminative feature as compared to color and shape. Jian and Wei (2010) proposed an SVM-based method to recognize three types of cucumber diseases.

Plant disease identification based on principal component analysis and neural networks (Wang, Li, Ma, & Li, 2012) was developed. They proposed a grape disease identification system in which principal component analysis was done to identify the unique feature and multilayer perceptron was used for classification. The dataset of grape diseases included downy mildew and powdery mildew. They had obtained the maximum recognition accuracy of 94.29%. Sannakki, Rajpurohit, Nargund, and Kulkarni (2013) came up with a method to identify two types of grape diseases. Segmentation was done using thresholding and anisotropic diffusion. K-means clustering was used to segment disease spots. The proposed method achieved better training accuracies when using hue features as compared to the saturation and intensity.

35.6 Disease identification using convolutional neural network

Disease identification using convolutional neural network (CNN) has achieved excellent results in recent years. A basic CNN is a sequence of three main layers, convolutional layer, pooling layer, and fully-connected layer as shown in Fig. 35.3.

For an input x to the i th convolution layer, the output is

$$y = \text{ReLU}(W_i * x) \quad (35.1)$$

where $W_i = [W_i^1, W_i^2, \dots, W_i^K]$ represents K number of filter kernels of the layer and $*$ denotes the convolution operation.

Each filter is an $M \times M \times N$ matrix where M is the window size of the filter and N is the number of input channels. The first few convolution layers extract low-level features (such as edges) from the input image. Additionally, nonlinearity is introduced at the convolution output through a Rectified Linear Unit (ReLU) function $\text{ReLU}(x) = \max(0, x)$. The output to each subsequent convolution layer is called an activation map. Pooling layer subsamples the activation map and hence allows position invariance of features in the input image. Among various types of pooling, max pooling has been chosen which allows downsampling by computing the maximum value of each filter window traversing the entire output matrix of the activation map. The deep network also consists of dropout layers to avoid overfitting. These layers are arranged after the pooling layer. Finally, the fully-connected layer is stacked at the top of last convolution

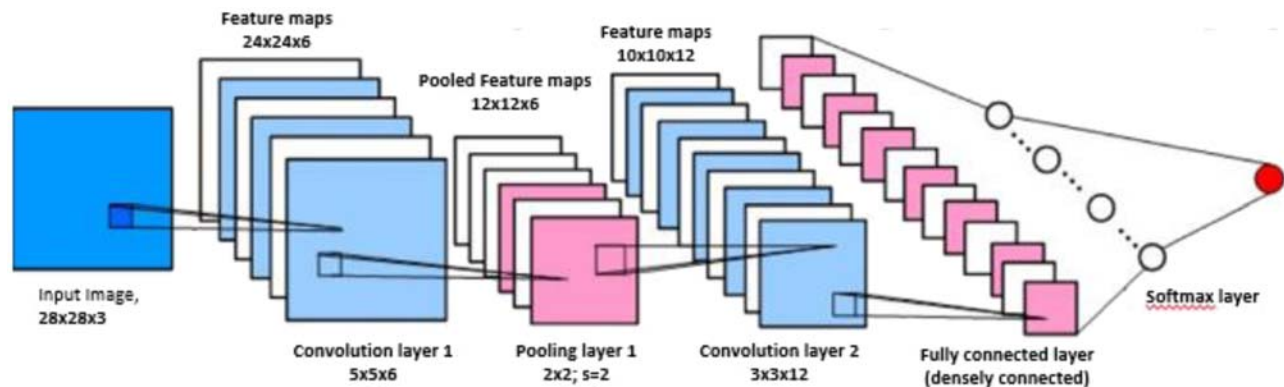


FIGURE 35.3 Basic structure of a convolution neural network.

layer. The last layer is the softmax layer which provides classification by exponentially normalizing the input which is fed from the last fully-connected layer. Arrangement of these layers in the form of a stack forms a CNN architecture. The loss function is used to measure the error between the predicted and the labeled input and is computed by

$$J = (W) = \frac{-1}{m} \sum_{i=1}^m y^i \log \hat{y}^i + (1 - y^i) \log(1 - \hat{y}^i) \quad (35.2)$$

where y is the expected and \hat{y} is the estimated output label vector, m is the number of training samples, w depicts the weight matrix of the convolutional and fully-connected layers.

The aim of training a network is to find an optimum value of w that minimizes the loss function, J . Backpropagation algorithms such as gradient descent and stochastic gradient descent are used to update weights.

Image-based plant disease detection using deep learning was proposed by [Mohanty, Hughes, and Salathé \(2016\)](#). They trained two deep learning models (AlexNet and GoogLeNet) using PlantVillage dataset. They had identified 26 diseases from 14 crop species. They examined the performance of the two deep neural networks by changing the training–testing distributions. They achieved a maximum accuracy of 99.35% using GoogLeNet when the training–testing distribution is made as 80–20.

An in-field automatic wheat disease diagnosis system was proposed by [J. Lu, Hu, Zhao, Mei, and Zhang \(2017\)](#). They developed a mobile application–based real-time wheat disease identification system. By implementing two different CNNs VGG-FCN-S and VGG-FCN-VF16, they obtained average recognition accuracies of 95.12% and 97.95%, respectively.

A deep learning–based real-time leaf disease’s detector for grapes using improved CNNs was proposed by [Xie et al. \(2020\)](#). They presented a deep learning–based Faster Dr-IACNN model with higher feature extraction capability for detecting grape leaf diseases.

35.7 Determination of the accuracy of the system

Accuracy of the system can be determined from the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values. A TP is an outcome where the model correctly identifies the positive class (healthy leaf image is identified as healthy by the system). Similarly, a TN is an outcome where the model correctly identifies the negative class (diseased image is identified as diseased leaf by the system). An FP is an outcome where the model incorrectly identifies the positive class (diseased image is identified as healthy). And an FN is an outcome where the model incorrectly identifies the negative class (healthy image is identified as diseased).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \quad (35.3)$$

To identify the accuracy of the system in each class (disease type), confusion matrix is used. Example of a confusion matrix is shown next.

There are 120 black measles present in the database, out of which the system predicts 115 images correctly and 5 images incorrectly. So for black measles class, the accuracy is 95.83%. Similarly for other classes, the accuracy can be calculated.

$$\text{Over all accuracy of the system} = 100 \times \left(\frac{115}{120} + \frac{100}{120} + \frac{120}{120} + \frac{50}{50} \right) = 94.79\%.$$

35.8 Severity estimation

Severity estimation plays a vital role in disease management. The main objective of severity estimation system is to calculate the amount of severity with which the plant has been infected. The severity estimation methods should match with the ground truth provided by the plant pathologist. Different parameters like fraction of infected area, rust color index number of lesions, number of epicenter, location of epicenter, texture of leaf, etc. are used to estimate the severity of the infection in a leaf image ([Cui, Zhang, Li, Hartman, & Zhao, 2010](#)). DNN can also be used to predict the severity in plants ([Wang, Sun, & Wang, 2017](#)). Fuzzy logic is also used to estimate the severity ([Negi & Tripathy, 2020](#)). The impressionness inherently present in the severity estimation process is handled by fuzzy inference system.

35.9 Conclusion

Different ML algorithms like SVM, K-means clustering, random forest are used to identify the diseases in plants. In these techniques, image database has to go through complex preprocessing steps like background removal, green channel enhancement, etc. Then the features are extracted from the preprocessed images. A neural network is obtained based on the healthy and diseased images. Once the training is done, when a new image is applied to the trained neural network, then the neural network can classify these images into a specific category. Even with complex image segmentation, feature extraction, and classification approach, these methods still have low disease identification accuracy. This happened because of the fact that low-level features (color, texture, and shape) fail to identify the high-level semantics (diseases in leaf). The CNN provides an end-to-end solution through deep learning. It takes the full advantage of image big data and identifies the discriminative features directly from the original image. To do so the CNN requires a large dataset. If the large dataset is not available, then image augmentation is done to increase the training and testing dataset. Because of the multilayer nature of the CNN, the computational cost is very high. So the need of the hour is to develop crop-specific lightweight CNN which can run on a smartphone and help the farmers to identify the disease at the earliest.

References

- Camargo, A., & Smith, J. S. (2009). Image pattern classification for the identification of disease causing agents in plants. *Computers and Electronics in Agriculture*, 66(2), 121–125.
- Cui, D., Zhang, Q., Li, M., Hartman, G. L., & Zhao, Y. (2010). Image processing methods for quantitatively detecting soybean rust from multispectral images. *Biosystems Engineering*, 107(3), 186–193.
- Dhaygude, S. B., & Kumbhar, N. P. (2013). Agricultural plant leaf disease detection using image processing. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2(1).
- Ghaiswat, S. N., & Arora, P. (2014). Detection and classification of plant leaf diseases using image processing techniques: a review. *International Journal of Advanced Engineering and Technology*, 2(3), 2347–2812.
- Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks/A Publication of the IEEE Neural Networks Council*, 13, 415–425.
- Jian, Z., & Wei, Z. (2010). Support vector machine for recognition of cucumber leaf diseases. In *2010 2nd international conference on advanced computer control*, IEEE, Shenyang, pp. 264–266.
- Lu, J., Hu, J., Zhao, G., Mei, F., & Zhang, C. (2017). An in-field automatic wheat disease diagnosis system. *Computers and Electronics in Agriculture*, 142, 369–379. Available from <https://doi.org/10.1016/j.compag.2017.09.012>.
- Meunkaewjinda, A., Kumsawat, P., Attakitmongcol, K., & Srikaew, A. (2008). Grape leaf disease detection from color imagery using hybrid intelligent system, In *2008 5th international conference on electrical engineering/electronics, computer, telecommunications and information technology*, IEEE, Krabi, pp. 513–516.
- Moghadam, P., Ward, D., Goan, E., Jayawardena, S., Sikka, P., & Hernandez, E. (2017). Plant disease detection using hyperspectral imaging. In *2017 international conference on digital image computing: techniques and applications (DICTA)*, Sydney, NSW, Australia, pp. 1–8. doi:10.1109/DICTA.2017.8227476.
- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419. Available from <https://doi.org/10.3389/fpls.2016.01419>.
- Negi, R., & Tripathy, S. S. (2020). Application of fuzzy logic in plant disease management, fuzzy expert systems and applications in agricultural diagnosis. *IGI Global*. Available from <https://doi.org/10.4018/978-1-5225-9175-7.ch013>.
- Rateria, A., Mohan, M., Mukhopadhyay, K., & Poddar, R. (2019). Investigation of *Puccinia triticina* contagion on wheat leaves using swept source optical coherence tomography. *Optik – International Journal for Light and Electron Optics*, 178, 932–937.
- Sannakki, S. S., Rajpurohit, V. S., Nargund, V. B., & Kulkarni, P. (2013). Diagnosis and classification of grape leaf diseases using neural networks. In *Proceedings of the fourth international conference on computing, communications and networking technologies*, Tiruchengode, pp. 1–5.
- Wang, G., Sun, Y., & Wang, J. (2017). Automatic image-based plant disease severity estimation using deep learning. *Computational Intelligence and Neuroscience*, Article ID 2917536.
- Wang, H., Li, G., Ma, Z., & Li, X. (2012). Image recognition of plant diseases based on principal component analysis and neural networks. *Proceedings of the 8th International Conference on Natural Computation, Okinawa Prefecture*, 246–251.
- Wijesinghe, R. E., et al. (2016). Optical inspection and morphological analysis of *Diospyros kaki* plant leaves for the detection of circular leaf spot disease. *Sensors*, 16, 1282.
- Xie, X., Ma, Y., Liu, B., He, J., Li, S., Wang, H., & Deep-Learning-Based, A. (2020). Real-time detector for grape leaf diseases using improved convolutional neural networks. *Frontiers in Plant Science*, 11, 751. Available from <https://doi.org/10.3389/fpls.2020.00751>.
- Yao, Q., Guan, Z., Zhou, Y., Tang, J., Hu, Y., & Yang, B. (2009). Application of support vector machine for detecting rice diseases using shape and color texture features. In *2009 international conference on engineering computation*, IEEE, Hong Kong, pp. 79–83.
- Youwen, T., Tianlai, L., Yan, N. (2008). The recognition of cucumber disease based on image processing and support vector machine. In *2008 congress on image and signal processing*, IEEE, Sanya, pp. 262–267.

Role of artificial intelligence, sensor technology, big data in agriculture: next-generation farming

Pradeep Kumar¹, Abhishek Singh², Vishnu D. Rajput³, Ajit Kumar Singh Yadav⁴, Pravin Kumar⁵, Anil Kumar Singh⁶ and Tatiana Minkina³

¹Department of Forestry, Applied Microbiology Laboratory, North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India, ²Department of Agricultural Biotechnology, College of Agriculture, Sardar Vallabhbhai Patel University of Agriculture and Technology, Meerut, Uttar Pradesh, India, ³Academy of Biology and Biotechnology, Southern Federal University, Rostov-on-Don, Russia, ⁴Department of Computer Science and Engineering, North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India, ⁵Department of Electrical Engineering, School of Engineering, Gautam Buddha University, Greater Noida, Uttar Pradesh, India, ⁶DeHaat, Agrevolution Pvt. Limited, Gurugram, Haryana, India

36.1 Introduction

Agriculture has traditionally been considered an intuitive place in which knowledge is passed from one generation to another. Nonetheless, currently, glitches such as the changing climate and lack of viable farming are major challenges. The United Nations estimates that the global population will reach 9.8 billion by 2050, an increase of 2.2 billion from this day and age. This means that to meet the demands due to the increasing number of people, there is a necessity to surge our crop production but unfortunately, hasty urbanization and climate change have windswept the major fragment of agriculture. In the United States alone, urbanization and climate change have worsened the total area of farms ranging from 913 million acres in 2014 to 899 million acres in 2018. In the current technology-based era the concept came into the picture which contributes to big data which refers to bulk number collection of soil architecture, weather forecast, fertilizer recommendation, disease management, climate change, crop mapping data (Manyica et al., 2019). These data are extracted by several resources like IoT (Internet of things) systems, software, and web portal (Nidhi, 2020). Nevertheless, big data information is implemented by robots and some forms of artificial intelligence (AI) (Fig. 36.1). By convention, farms have looked for many workers, customarily seasonal, to harvest crops and keep farms productive. However, society has moved away from being an agrarian society with large quantities of people living on farms to people living in cities now; thus farms are facing the challenge of a workforce shortage. One way to aid with this shortage of workers is agricultural robot integrating AI features. According to a Forbes study (Walch, 2020), farm robots augment the human labor workforce and can harvest crops at a higher volume and faster pace than human (Saiz-Rubio and Rovira-Más, 2020). Although there are still many cases in which robots are not as fast as humans, agriculture is currently developing robotic systems to work in the field and help producers with tedious tasks (Saiz-Rubio & Rovira-Más, 2020), pushing agricultural systems to the new concept of agriculture. According to Reddy, Reddy, and Kumar (2016), the advent of robots in agriculture drastically increased productivity in several countries and reduced farm operating costs. As said earlier, robotic applications for agriculture are mounting exponentially (Shamshiri et al., 2018), which bids promising solutions for smart farming handling labor shortage and a longtime declining profitability. However, like most innovations, there exist important limitations to cope with the current initial stages. These technologies are still too exclusive for most farmers, especially those owning minor farms (Lamborelle & Fernández Álvarez, 2019) because scale economics make small individual farms less profitable (Sonka, 2014). Nevertheless, the cost of technology shrinks with time, and agricultural robots will be surely be implemented in the future as an alternative to bring about sophisticated production (Saiz-Rubio & Rovira-Más, 2020). The world's

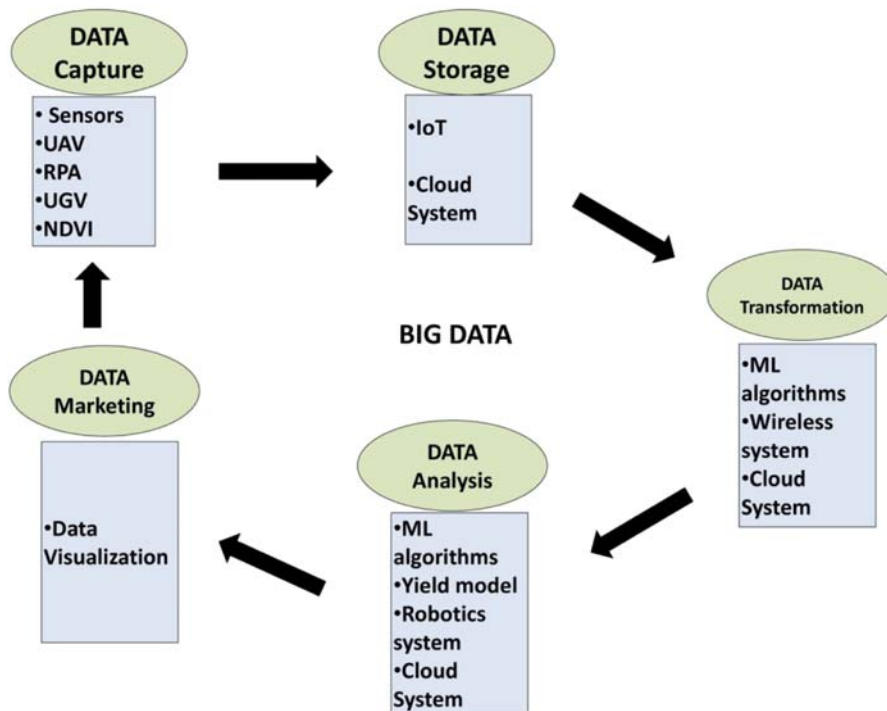


FIGURE 36.1 Diagrammatic representation of big data information cycle.

agricultural production and crop yield slackened down in 2015. The notion of agricultural robotics was introduced to overcome these problems and satisfy the rising demand for high yields. Robotic innovations are giving a boost to the global agriculture and crop production market, also according to the Verified Market Intelligence report, agricultural robots will be capable of completing field tasks with greater efficiency as compared to the farmers (Verified Market Intelligence, 2018). Agricultural tech startups have upstretched over 800 million dollars in the last 5 years (Cbinsights, 2019). Startups consuming robotics and machine learning (ML) to disentangle snags in agriculture started gaining momentum in 2014, harmonized with an intensifying interest in AI (Varadharajan, 2019). Venture capital funding in AI has increased by 450% in the last 5 years (Murugesan et al., 2019). This kind of new agriculture pretends to do additional with less, for the reason that nourishing people while increasing production sustainably and taking care of the environment will be decisive in the coming years, as the Food and Agriculture Organization of the United Nations estimates that, in 2050, there will be a world population of 9.6 billion (Zhang, 2015). Advanced sensing technologies in agriculture can help to meet the challenge; they provide detailed information on soil, crop status, and environmental conditions to allow precise applications of phytosanitary products, resulting in reduced use of herbicides and pesticides, amended water use efficiency, and increased crop yield and quality (Zhang, 2019).

36.2 Characteristics of big data

Big data is metaphorically compared to the ocean and information to a floating Iceland inside it which can be accessed by different platforms and software. Big data can be pigeonholed by four Vs—volume, velocity, variety, and veracity.

36.2.1 Volume

The volume raises the huge amount of data that are captured, stored, managed, and analyzed through an eclectic range of resources.

36.2.2 Velocity

Velocity refers to the data that need to be collected, stored, processed, analyzed in real time, for example, Google Maps that provided us real-time data of traffic, soil, weather, etc.

36.2.3 Variety

A variety of data refer to the data arising from multiple formats such as structured data in the traditional database and unstructured text documents such as email, video, audio, and financial transactions.

36.2.4 Veracity

Veracity is another term that characterizes big data signifying access to heaps of data, generated from assorted sources with minimal lag times and data flows that vary greatly with periodic peaks and baseline.

36.3 Big data and smart agriculture

Big data and smart agriculture both are reasonably newfangled concepts of agriculture. The precision agriculture concept is an extension of smart agriculture; based on big data information, the farmer takes the decision and accomplishes the situation as per information provided by big data. The big data contain a key feature that is real-time assistance like suddenly changed operational conditions or other circumstances, for example, weather or disease alert, through these real-time assistance crop management systems like weather alert system, crop sensor, pest spraying UAV (unmanned aerial vehicle) carry out agile actions (Esmeyjer, Bakker, Ooms, & Kotterink, 2015). Big data also have intelligent assistance which helps in the implementation, maintenance, and use of agriculture technology as well. The main role and application of big data in smart farming are to ensure minimum cost gaining higher profit as well as sustainability. The use of AI, sensors, and smart machines in agriculture has brought agriculture to the top of the digital revolution in the modern era. Data in agriculture are a collection of data about various types of soil mapping containing information related to their physical and chemical properties, weather, past management practices, etc., because of all this data information, in any adverse weather or diseases, the farmer is warned in advance, due to which the farmers agonize less (Nidhi, 2020). Big data for smart agriculture contain a far-reaching level collection of good agriculture practices data, these are the following.

36.3.1 Digital soil and crop mapping

Digital soil and crop mapping concern with building digital maps for soil types and their physiochemical properties. In developed and some developing countries, farmers supervise so many acres of land, it is almost difficult to get instant updates and alerts from their planted lands without exhausting technology. Aimed at the management and inspection of these many acres of the lands, many countries like Ireland use satellite-based soil and crop monitoring that are more rapid and cost-effective than traditional methods.

36.3.2 Weather prediction

Crops growths, development yield, and total agriculture production hinge on weather. Not only in India, but also in many other countries like Bangladesh, Pakistan, Japan, Korea, China, agriculture system is also influenced by the weather. In other words, India and other countries' agriculture system is weather based. Weather aberrations can ground physical damage to crops and soil erosion. Abrupt weather changes source severe damage to crops. All economical processes, including the quality of crops from agricultural land to market, transportation storage, depend on the weather. Debaunched weather hampers all aspects mentioned previously ensuing in high economical losses. Agricultural weather conjectures the following elements:

1. Low-pressure areas, cyclones, tornadoes, and depressions
2. Wind speed and direction
3. Relative humidity
4. Max, min, and dew point temperatures
5. Amount and type of coverage of sky by clouds
6. Rainfall and snow
7. Events like fog, frost, hail, thunderstorms, and wind squalls
8. Earth observation satellites, UAV (Drones), and automatic weather stations

36.3.3 Fertilizers recommendation

Eloquent exact fertilizer rate for crop field is a science and this science requires analysis of multiple factors, parameters at the nanolevel. These parameters comprise crop nutrient uptake rates; research data; soil chemical, physical, and biological properties; weather; water composition; land type; soil testing methods; irrigation techniques; fertilizer characteristics; interactions of fertilizers between crops. The use of this excess amount of fertilizer in agriculture field into the soil as toxic compounds gives rise to various types of pollutants. Big data tools are now able to advise the farmers with the right quantity of fertilizers.

36.3.4 Disease detection and pest management

In modern agriculture, with the help of big data tools, developed advanced algorithms are used to identify the patterns and behavior of various types of microorganisms and pests which helps in forecasting the invasion of pests and the spread of microscopic diseases. Agricultural pests can quickly censor into a farmer's revenues, but misusing and a higher amount of pesticide use can have adverse effects on people (like they can cause cancer), plants, and other living things. UAV and crop sensors assist in pest control, mid-season crop health monitoring reducing the use of pesticides.

36.3.5 Adaptation to climate change

Climate change due to global warming is a looming concern that affected the agriculture sector. One project of big data provides IoT sensors to Taiwanese farmers for rice production so that they can assemble information that is necessary about their crops. The collective information of IoT sensors will help farmers to optimize their production cycles even in antagonistic climate conditions. Traditional farming is not able to analyze these climate changes due to which traditional farmer faces gigantic economical loss. With the solution to all these concerns, big data can revolutionize the future of farming.

36.3.6 Automated irrigation system

All the countries in the world are currently in a situation where they are required to use water in a very resourceful manner. According to recent studies, water is flatter more and more in short supply worldwide and over one-third of the world population would aspect total water shortage by the year 2025. In agriculture as well, the major problem which farmers face is water scarcity, hence to improve the convention of water, one of the irrigation systems—using drip irrigation which is implemented as an automated irrigation system for small-scale farms, and the other being automated irrigation system using weather prediction.

36.4 Sources of big data

Large scale and a wide variety of sources can originate big data. Sources of big data stand to be ground sensors (chemical detection devices, biosensors, weather stations, etc.) which observed that the farmers filed and provided the data, governmental organizations, NGOs, and other private organizations collect the data (statistical yearbook, government) which is also a major source for big data (Fig. 36.2) (Chedad, 2001; Kempenaar, Lokhorst, Bleumer, & Veerkamp, 2016). Online stored and web service data obtained from airborne sensors like UAV, light airplanes, satellites are a source for big data (Becker-Reshef, 2010; Gutiérrez, 2008). Cloud system that is an amalgamation of wireless sensor networking, IoT system source for big data, provides real-time web data (information about plants, crops, yields, weather conditions, etc.) from private companies to farmers on their mobile phones; media are (social media platforms corresponding Facebook, Twitter, YouTube, Instagram agriculture channels, videos, audios, research reports, articles) also sources for big data. Altered sources of big data are required to deal with the various problems in agriculture and this big data is used in agriculture applications. For example, crop-, soil-, and animal-related research use ground sensors deployed at the field, climate change applications use data from weather stations, information on land mapping from satellite, data and this information used by many government agencies to making the policies for farmers to enhance production, export and import predication as per yield of crops. Various types of government and private service providers have collected the agricultural big data information for its service to the consumer (Table 36.1).

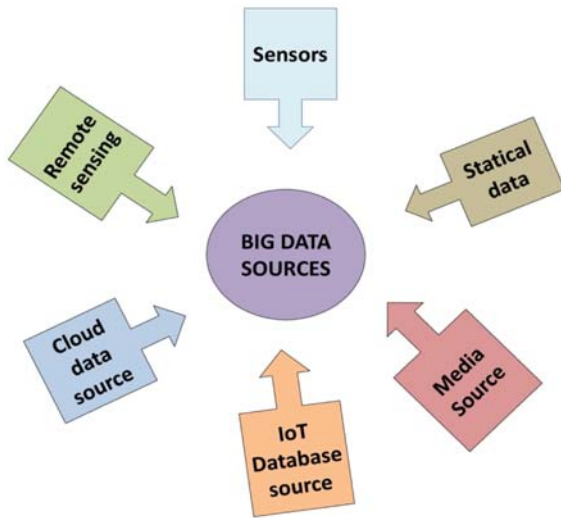


FIGURE 36.2 Diagrammatic representation of big data sources.

TABLE 36.1 Overview of big data sources, service providers.

Agriculture area	Big data source	Big data innovative service providers
Weather and climate	Earth observation satellites UAV, automatic weather stations	Fertilizer calculator Crop water needs Estimation Yield models
Soil	Mobile app’s location-based datasets ground sensors/base stations, UAV	Soil indicators for Scottish soils, SoilInfo, SoilWeb mKrishi (India), CiAgriculture (China), Fujitsu (Japan) Batian (China), RedBird, etc.
Crop and yield	Fertilizer calculator Crop water needs Estimation Yield models	BaiKhao

Source: From Swiss Re Centre for Global Dialogue Switzerland.

36.4.1 Sensors

Sensors are the radical devices that monitor crops and obtain objective information from them. Sensors can be classified according to their platforms, like satellites, aerial, or airborne (airplanes, UAVs, balloon), and ground based. Airplanes, satellites, and UAVs mostly employ cameras to amass images and ground-based optical sensors that can collect reflectance data and storage in a text file (Saiz-Rubio & Rovira-Más, 2020) (Fig. 36.3).

36.4.1.1 Remote sensing platforms: satellites

Remote sensing has frolicked a crucial role in collecting the data related to geographical characteristics of a particular area and charting these geographical areas without physical contact with the areas that have to be measured with the images gathered from the satellites (Ma et al., 2015). In the world, every country is hurling satellites for collecting real-time data allied to agriculture. America has eight Landsat satellites that gross spectral data from the Earth each 16–18 days; European Sentinel has two Landsat satellites that provide multispectral data at 10-m pixel resolution for NDVI—Normalized Difference Vegetation Index—imagery, soil, and water cover in every 10 days; additionally, the RapidEye being a German geospatial information provider operated a five-satellite constellation and provided multispectral RGB imagery, as well as red-edge and near-infrared (NIR) bands at 5-m resolution; GeoEye-1 being another satellite had been launched on September 6, 2008 to capture multispectral RGB data and NIR data at a 1.84-m resolution;

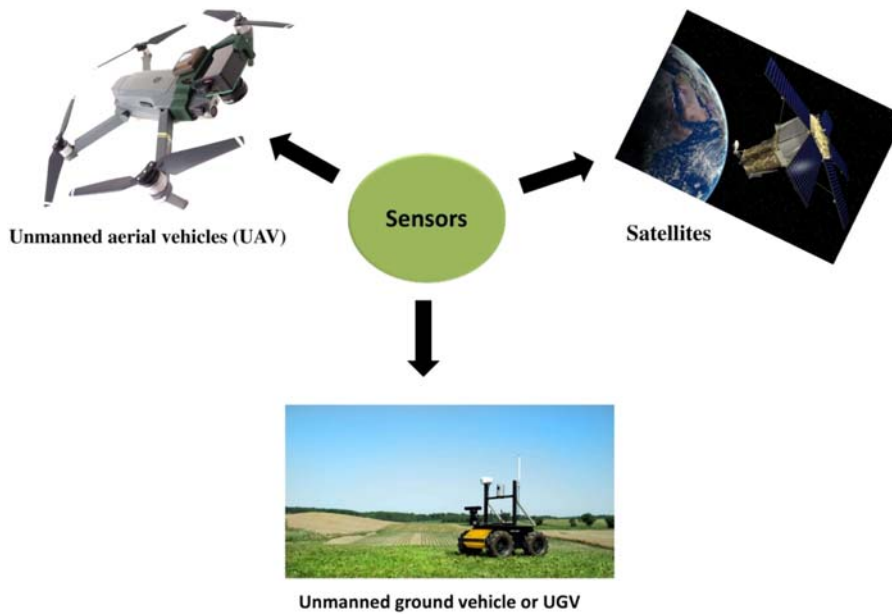


FIGURE 36.3 Diagrammatic representation of sensors.

WorldView-3 satellite launched on August 13, 2014 collects multispectral data from the RGB bands, including the red-edge, two NIR bands, and eight SWIR bands with a resolution of 1.24 m at the nadir (Saiz-Rubio & Rovira-Más, 2020). Several studies focus on the potential applications of thermal technologies using remote sensing and determine the nutritional status of field crops (Khanal, Fulton, & Shearer, 2017; Rudd, Roberson, & Classen, 2017).

36.4.1.2 Airborne platform systems: unmanned aerial vehicles and remotely piloted aircraft

The distance between a satellite and a crop is very extraordinary ranging up to 700 km and deeper insights are reachable when sensors endure closer to the targets. The distance between the airborne system and the land is about 100 m so that it can easily inspect the land and get information from there about the crops, soil, fertilizer, crop disease, etc. UAVs and remotely piloted aircraft are mainly divided into two types, fixed-wing aircraft and multirotor aircraft. Fixed-wing UAVs can cover more agricultural area per flight and carry larger payloads but easily break after multiple landings and are supplementarily expensive (Rudd et al., 2017). Rotary-wing UAVs are more stable fliers as they are capable of vertical takeoff and landing; however, they are slower and cannot conceal as much area during their battery life (Saiz-Rubio & Rovira-Más, 2020). UAVs have more advantages than remote sensing, which include frequency flexibility and better spatial resolutions that help to collect big data in agriculture. Compared to ground vehicles, UAVs can obtain data from inaccessible locations where conventional equipment cannot stance. Nevertheless, they require advanced professional planning of the flight path and some machine vision applications that may need to soar in the afternoon to avoid vegetation shadows on the ground, causing errors with imagery data (Rudd et al., 2017).

36.4.1.3 Ground platform systems: unmanned ground vehicle

Ground platform systems' sensors are known as unmanned ground vehicles or UGVs. This UGV is equipped with advanced technologies for positioning and orientation, navigation, planning, and sensing which are applied in agriculture for tilling, soil analysis, seeding, transplanting, crop scouting, pest control, weed removal, and harvesting (Bechar & Vigneault, 2016; Bechar & Vigneault, 2017; Corke, Roberts, & Winstanley, 1998). Husky is a medium-sized robotic UGV developed by Clearpath Robotics company which has a role to carry some unique features like stereo cameras, LIDAR, GPS, IMUs, manipulators that can be beneficial for monitoring field crops and collected data.

36.4.2 Statistical data

A massive level of agricultural data is prepared by governmental organizations, NGOs, and other private organizations which are collected in the form of the data and published, for example, statistical yearbook, survey data of government-

related weather, soil properties type, soil mapping, crop mapping, water-related data, etc. which stamp as a major source for big data (Chedad, 2001; Kempenaar et al., 2016).

36.4.3 Remote sensing

Remote sensing has played a crucial role in collecting the data related to geographical characteristics of a particular area and they chart these geographical areas without physical contact with the areas that have to be measured with the images congregated from the satellites (Ma et al., 2015). In remote sensing Earth the satellites collect data allied to crop mapping, soil mapping, water availability which also forecast drought and flood (Shelestov, Lavreniuk, Kussul, Novikov, & Skakun, 2017).

36.4.4 Cloud data source

Cloud computing is emerging today as a saleable infrastructure associated with a new paradigm for the provision of computing infrastructure and big data processing methods for various resources (Patel & Patel, 2013). It eliminates the need for maintaining expensive computing hardware, software, Information technology, staff, infrastructure, recourses, and their maintenance. Various types of digital tools in agriculture like sensors, remote sensing, UAV, etc. provided the data bank related to soil, weather, crop, farmers, agriculture marketing, fertilizers, and pesticide information which are hoarded in a single place in the cloud. This big data from the cloud can be easily accessed by the end users such as farmers, experts, consultants, researchers with the help of various software in the form of mobile applications, web portals, etc.

36.4.5 Internet of things database source

IoT is a new era of revolutionary technology that empowers to connect the object (such as plants) and devices (such as sensors) to enormous databases via the help of the Internet. IoT facilitates interaction between agricultural objects. Various agricultural devices provided output data in variable formats. Hence it is very important to compile all these data in a common protocol for communication between the objects and devices in the network. After one-to-one care of the crops during the production, it is also required that the agricultural products have to be tracked after harvest; radio-frequency identification (RFID), wireless sensors networks (WSN) serve as a basic building block for using IoT in agriculture, WSN plays a strategic role in intensive care of the storage and logistics facilities of the yield (Madhuri & Indiramma, 2019). These WSN sensors can be deployed to measure soil pH, soil moisture, water, fertilizer, pest, temperature, evaporation (Yan-e, 2011). These different kinds of sensors collect the data to sense the growth patterns changes in plants like plant height measurements, chlorophyll measurements, leaves area index, rate evaporation, etc. through capturing the images of plants through RFID to track all these patterns (Zhao, Zhang, Feng, & Guo, 2010). The big data information collected by IoT-based sources help the farmers to better understand plant health, plant height measurements, chlorophyll measurements, weed pressure charting for microlevel management to the application of irrigation pesticide herbicide and fertilizer (Madhuri & Indiramma, 2019).

36.4.6 Media source

Media corresponding to social media platforms, for example, Facebook, Twitter, YouTube, Instagram agriculture channel, videos, audios, research reports, articles are also a source for big data.

36.5 Techniques and tool use in big data analysis

Big data are highly dimensional and heterogeneous, containing complex information, sophisticated tools, and techniques that are needed to extract information from it (Mucherino, Papajorgji, & Pardalos, 2009; Vitolo, Elkhatib, Reusser, Macleod, & Buytaert, 2015). Big data are analyzed by using various algorithms, single or combination of two different techniques, for example, ML image processing, remote sensing, cloud platforms, GIS (geographic information system), vegetation indices (VIs), NDVI (Kamilaris, Kartakoullis, & Prenafeta Boldú, 2017).

36.5.1 Machine learning

ML is a prevalent technology nowadays that can be used in agriculture for more sustainability. The usage of ML in agriculture makes it more lucrative because of its micro-label management. ML also contains artificial ML which is a bottleneck in applied agricultural science. Artificial techniques are being used in the agricultural sectors to amplify accuracy and to seek solutions to the hitches.

36.5.1.1 Livestock management

The livestock is categorized into two parts, animal livestock and livestock production. Animal welfare deals with animal health, welfare, disease, and well-beings. The main application of ML is monitoring the early detection of animal diseases and behavior. It also scans the economic profit as well as losses and production of animal-related goods.

36.5.1.1.1 Livestock production

Adhering to human civilization, domesticated livestock has played a fundamental role henceforth becoming an integral part of human culture, society, and, most importantly, the global economy. Domestic livestock has underwritten the rise of human societies and civilizations by cumulating the amount of food and nutrition available to people in four ways: by providing sources of meat, milk, and fertilizer and by pulling plows. Throughout antiquity, livestock has also provided leather, wool, other raw materials, and transport. With the help of ML analysis the accurate data prediction of rumen fermentation suggested the diets, fatty acid percent in milk, and milk production of animals (Craninx, Fievez, Vlaeminck, & De Baets, 2008). ML model support vector machines are useful for early uncovering and warning of problems in the commercial production of eggs, which helps the poultry industry to analyze data related to hen and egg production (Mohammadi et al., 2015). ML algorithm-based convolutional neural networks are effectively applied in digital images for face recognition of the animal, for example, pig and this method overcomes the problem of distressing activity for tagging of RFID which has a lot of limitations like low range and time-consuming (Hansen et al., 2018).

36.5.1.1.2 Animal welfare

Animal welfare is defined as the relationships of humans with animals and the duty they pledge humane and responsible treatment to the animals under their care. Animals collar sensors with magnetometers and three-axis accelerometers data are collected and analyzed by ML modeling preceded by the events such as the estrus and the recognition of dietary changes in animals (Dutta et al., 2015). ML acts on chewing patterns data with a combination of behavioral data of calves like dietary supplements, such as hay and ryegrass, which were collected by optical FBG sensors analyzed by ML (Pegorini et al., 2015).

36.5.1.2 Water management

In the agriculture system, water is a fundamental aspect of farming showing the importance of water in farming so it requires micromanagement and plays a significant role in hydrological, climatological, and agronomical balance. ML is used in the estimation of daily, weekly, or monthly evapotranspiration (Mohammadi et al., 2015). The progression of evaporation is very complex for every plant, but subsequently understanding this process, the irrigation system can be developed to provide water according to the requirement of each plant in the crop field. This is also important in the present 2020 scenario because according to recent studies, with continuous use of groundwater, more than one-third of the world population would face total water shortage by the year 2025. Temperature sensors provided data which were analyzed by ELM model of ML to estimate the accurate weekly evapotranspiration in the arid region of India; therefore this big data analyzing model would help in crop water management (Patil & Deka, 2016).

36.5.1.3 Soil management

Soils are the superlative media for the growth and development of each plant. For healthy growth of plants, it requires healthy soil and ML application to predict identification and estimate the soil properties like soil drying, condition, temperature, and moisture content (MC). Earth has oodles of biodiversity climate and geographical distribution of lands that is why the soil of each in every area is a heterogeneous natural resource, with complex processes and mechanisms that are difficult to understand (Johann, de Araújo, Delalibera, & Hirakawa, 2016). Combination of ML and big data related to the soil developed a method for the provision of remote agricultural management decisions, which evaluated

soil drying for agricultural planning. This smart method accurately gauges the soil drying, with evapotranspiration and precipitation data, in a region located in Urbana, IL of the United States (Coopersmith, Minsker, Wenzel, & Gilmore, 2014). Big data are also a master collection of soil condition data such as composition, nutrient availability, mineral quantity, and type of soil. Various modeling aspects of ML with combination soil data predicated soil organic carbon, MC, and total nitrogen, for this analysis used a visible–NIR spectrophotometer to collect soil spectra from 140 unprocessed and wet samples of the top layer of Luvisol soil types, which were collected from an arable field in Premslin, Germany in August 2013, after the harvest of wheat crops (Morellos et al., 2016). Soil moisture is also estimated by ML artificial neural network model and data are obtained from force sensors (Johann et al., 2016).

36.5.2 Cloud platforms

Cloud computing is an information technology paradigm through which users can access shared pools of configurable system resources over the Internet. Cloud platforms unruffled with big data sources (crop, weather and climate, soil, growth progress, and pattern) collected data that need to be accumulated at a common platform that should be easily accessible, preprocessed, visualized, and analyzed (Kamilaris et al., 2017). Cloud computing of agriculture is used to manage to analyze the environmental factors, analyze the soil moisture, temperature, and manage the water supply functions (Murakami, Utomo, Hosono, Umezawa, & Osawa, 2013). Cloud system allows farmers to view farm or farm field information with ground sensors, devices connected, etc. Apart from this, the system allows farmers to control the farm hardware remotely such as to switch on/off bulb and motors with the help of microcontroller (Balbudhe, Amar, Nikhil, Saket, & Nandan, 2015). Cloud system has the following role in the case of big data:

1. Cloud system stores all the agriculture-related information provided by big data sources in a centralized cloud, which will be available to all the users like a farmer, agricultural companies, etc., at anytime, anywhere.
2. Cloud system management of all big data is related to land, location, area; soil, and land characteristics through centralized decision support systems.

36.5.3 Geographic information systems

GISs are computer hardware and software that use feature attributes and location data to produce maps (Lucas & Chhajed, 2004). GIS has imperative feature functions in agriculture like storing layers of information, such as yields, soil survey maps, remotely sensed data, crop scouting reports, and soil nutrient levels (Barrett, Nitze, Green, & Cawkwell, 2014). GIS tool combined with big data source provided by Earth observation satellites, UAV, ground sensors like temperature sensors, moisture sensors is collected as large volume of data which observed by GIS analyzing software in various forms like visual image, audio, video for weather forecasting (Kamilaris et al., 2017).

36.5.4 Vegetation indices

UAV pooled with a thermal camera and satellite sensors provided data related to measuring wavelengths of light absorbed and reflected by green plants. In the biological system, every plant leaves contain a pigment system that strongly absorbs wavelengths of visible (red) light and, on the other hand, strongly reflects wavelengths of NIR light, which is invisible to human eyes. As a plant canopy changes from early spring growth to late-season maturity and senescence, these reflectance properties also change. VIs obtained from remote sensing-based canopies are effective algorithms for quantitative and qualitative evaluations of vigor, vegetation cover, and growth dynamics, among other applications (Xue & Su, 2017). VIs are various types like NDVI, GDVI (Green Normalized Difference Vegetation Index), and SAVI (Soil Adjusted Vegetation Index) but NDVI is widely used to obtain and analyze the big data in the agriculture. In a biological system to govern the density of green plants on lands, it must observe the distinct colors (wavelengths) of visible and NIR sunlight reflected by the plants (Skakun, Justice, Vermote, & Roger, 2018). If sunlight passes the prism then many different colors of light with different wavelengths called VIBGYOR (Violet–Indigo–Blue–Green–Yellow–Orange–Red) are visible and when sunlight strikes objects, certain wavelengths of this spectrum are absorbed and other wavelengths are reflected. The pigment in plant leaves, chlorophyll, strongly absorbs visible light (from 0.4 to 0.7 μm) for use in photosynthesis. The cell structure of the leaves, on the other hand, strongly reflects NIR light (from 0.7 to 1.1 μm) (Daroya & Ramos, 2017). As many leaves a plant has, there is an increase in effect to these wavelengths of light, respectively. NDVI values range from +1.0 to –1.0. Areas of barren rock, sand, or snow usually show very low NDVI values (e.g., 0.1 or less) (Mahajan & Raj, 2016). Sparse vegetation such as shrubs and grasslands or senescing crops may result in moderate NDVI values (approximately 0.2–0.5).

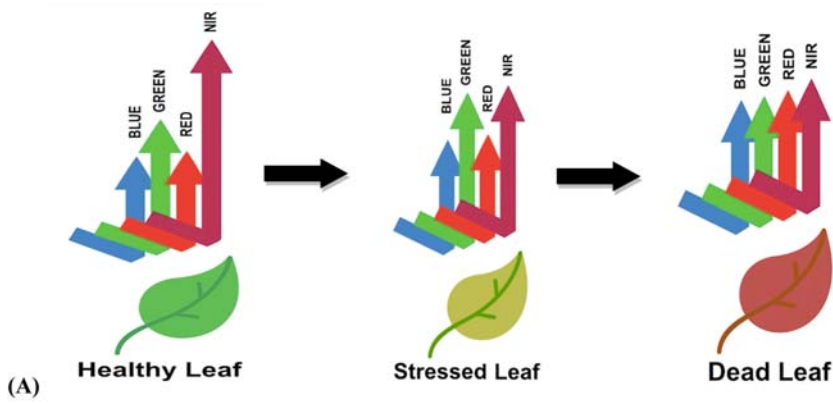


FIGURE 36.4 (A and B) Diagrammatic representation of NDVI. *NDVI*, Normalized Difference Vegetation Index.

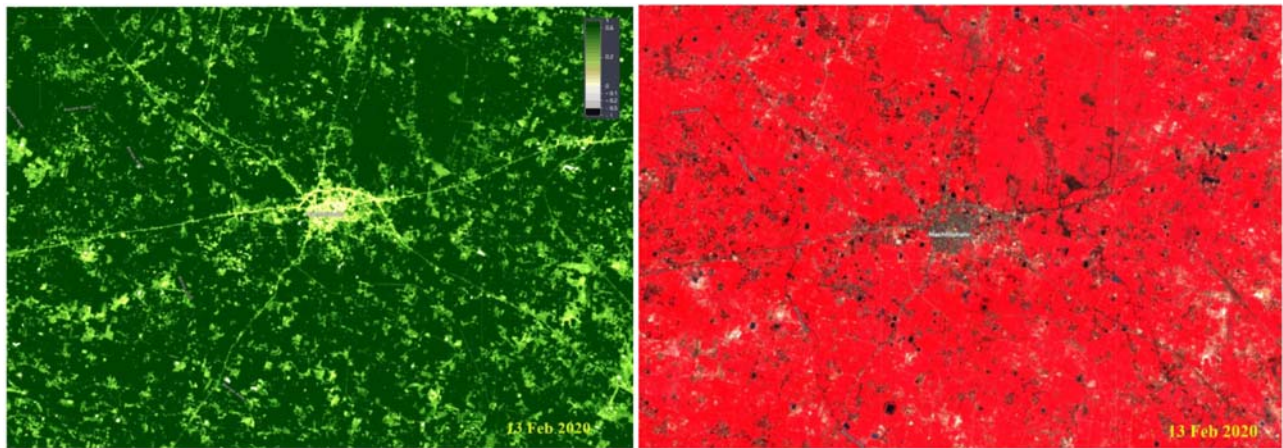
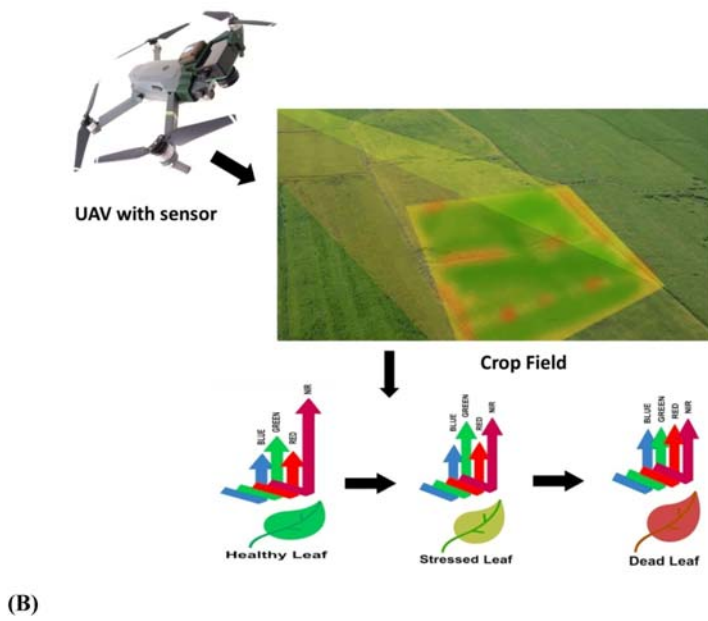


FIGURE 36.5 Satellite images of Machhlishahr, Jaunpur, (A) NDVI (B) FCC. *FCC*, False-color composite; *NDVI*, Normalized Difference Vegetation Index.

TABLE 36.2 Small overview of big data source, analyzer tools, software, and service provider companies.

S. no.	Agriculture field	Big data source	Big data analyzer tools	Software	Service providers
1.	Weather and climate	Weather stations, surveys, static historical information (weather and climate data, Earth observation data), remote sensing (satellites), geospatial data	Machine learning modeling algorithms, image processing, remote sensing, cloud platforms, GIS	Weather and Climate Toolkit (WCT), ArcGIS, Google Earth, MatLAB, QGIS, ModelVis	Weather Decision Technologies, Inc (WDT), IBM The Weather Company, CropProphet Enterprise, World Climate Service, AccuWeather, AerisWeather, Atmograph, Hurricane Mapping
2.	Soil management	Mobile app's location-based datasets ground sensors/base stations, UAV	Machine learning modeling algorithms, GIS, remote sensing, NDVI vegetation indices	SOILMAP, Soil Data Viewer 5.1, SOTER ArcGIS, Autodesk	Soil indicators for Scottish Soils, SoilInfo, SoilWeb mKrishi (India), CiAgriculture (China), Fujitsu (Japan) Batian (China), RedBird
3.	Crop and yield	Fertilizer calculator crop water needs estimation yield models	Machine learning modeling algorithms ANN, VIS-NIR, GIS, remote sensing, NDVI vegetation indices	AgroMetShell, SCOPUS, BaiKhaoNK, SOCiT, PocketLAI, RaGPS, MapIT, m-Sahayak, mKRISHI	Soil indicators for Scottish soils, SIFSS, ICAR, IASRI, BaiKhao
4.	Animal research and management	Livestock production	Machine learning modeling algorithms SVM CNNs, RFID, face-detection sensors	Ranch Manager, Chetu, Cattle Max, Livestock	SCR dairy, Lely Qwes

ANN, Artificial neural networks; CNNs, convolutional neural networks; GIS, geographic information system; NDVI, Normalized Difference Vegetation Index; RFID, radio-frequency identifications; SVM, support vector machines.

High NDVI values (approximately 0.6–0.9) correspond to dense vegetation such as that established in temperate and tropical forests or crops at their peak growth stage (Shafi et al., 2019). NDVI is a key tool to obtain and analyze big data related to crop health. All sensors like satellite, UAV collected broad data analyzed by smearing a different mathematical formula to quantify the density of plant growth on the Earth NIR radiation minus visible radiation divided by NIR radiation plus visible radiation obtained resulted from this mathematical calculation called NDVI (Shafi et al., 2019) (Figs. 36.4 and 36.5; Table 36.2).

$$NDVI = (NIR - VIS) / (NIR + VIS)$$

36.6 Role of big data in agriculture ecosystem: for smart farming

Keeping in view the big data source and big data analyzed tool, it has been found that with the help of big data, common farming will be converted into smart farming resulting farmers can earn good profits at low-cost wages (Wheeler & von Braun, 2013). For the production of the crop in a sustainable manner, it is required to utilize agricultural resources in a more precise way and in time, a decision for maximum resource utilization. Big data play important role in different aspects for smart farming like GPS for mapping, navigation, and IoT connected to remote sensors and monitoring system base autonomous driverless tractor (Conesa-Muñoz, Gonzalez-de-Soto, Gonzalez-de-Santos, & Ribeiro, 2015; Reeve, Eizad, & Ramm, 2011); smart autonomous machines and robotics base seedbed preparation to reseeding (Blackmore, Stout, Wang, & Runov, 2005; Griepentrog, Nrremark, Nielsen, & Blackmore, 2005); helicopters base smart planting from air to field is cost-effective for larger size of lands which is managed by GPS and IoT system (Pedersen, Fountas, & Blackmore, 2008; Scott, 2010); digital management of planting and

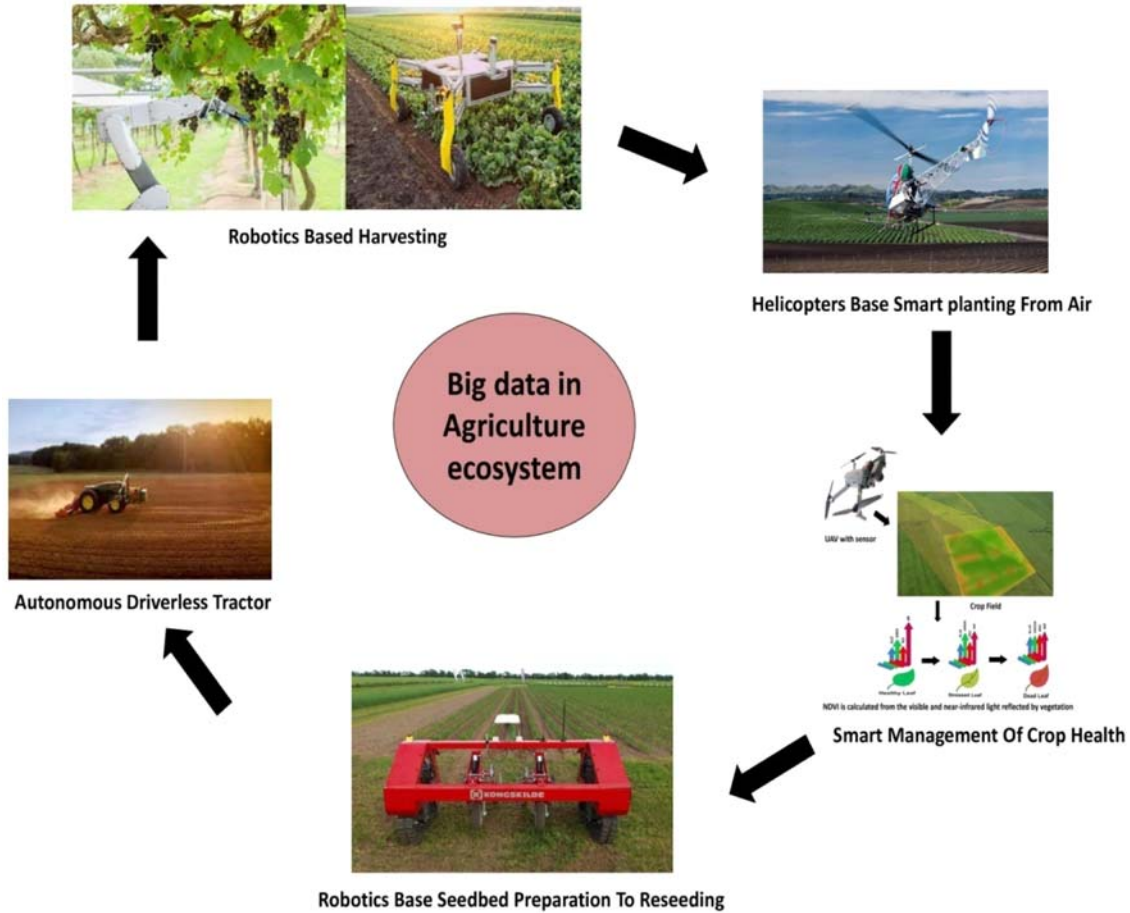


FIGURE 36.6 Diagrammatic representation the role of big data in agriculture ecosystem.

sowing through automation and robotics machine monitored by IoT and cloud system (Henten, Van, Bac, Hemming, & Edan, 2013; Buning, 2010); smart management of crop health through ML and AI, both integrated and applied to make it easy for farmers for the detection of pest, weed management, and crop health through image processing like NDVI, along with all the information, UAV, RPA, UGV technology is also widely adopted in many smart farms for spraying of herbicides, pesticides, fertilizer, and weather broadcasting (Veroustraete, 2015; Alimuzzaman, 2016). Crop yield analysis with the help of crop mapping through satellite UAV, NDVI is preparing data related to crop production which is stored in the cloud platform that estimates the yield of a specific location (McBratney & Whelan, 1999; Luck & Fulton, 2015). Smart method of harvesting from field robotics-based harvesting has an aim like the efficient ability to analyze the maturity of crops and harvest it without damaging the grains. The robotic system has a sensor that analyzes the ripening and maturing of fruit and crops before the harvesting ensuring that fruit and crops are perfectly ripened and mature last they start to harvest it (Blackmore et al., 2005; Yamamoto et al., 2010; Hayashi et al., 2011) (Fig. 36.6).

Acknowledgments

All authors are highly grateful to the authority of the respective department and institution for their support in conducting this research.

Conflict of interest

The authors declare no conflict of interests.

Author contributions

P.K. conceived and designed the manuscript; P.K., Ab. S., V.D.R., A.K.S.Y., P. Ku., A. S., and G.S. wrote the manuscript; P.K. and V.D.R. critically reviewed the manuscript and did the required editing.

References

- Alimuzzaman, M. (2016). *Agricultural drone*. Available from <https://doi.org/10.13140/RG.2.1.1146.2247>.
- Balbudhe, K. S., Prof. Amar, B., Nikhil, D., Saket, R., & Nandan, J. (2015). Cloud based cultivation management system. *ACSIJ Advances in Computer Science: An International Journal*, 4(3), No.15.
- Barrett, B., Nitze, I., Green, S., & Cawkwell, F. (2014). Assessment of multi-temporal, multi-sensor radar and ancillary spatial data for grasslands monitoring in Ireland using machine learning approaches. *Remote Sensing of Environment*, 152(2), 109–124.
- Bechar, A., & Vigneault, C. (2016). Agricultural robots for field operations: Concepts and components. *Biosystems Engineering*, 149, 94–111.
- Bechar, A., & Vigneault, C. (2017). Agricultural robots for field operations. Part 2: Operations and systems. *Biosystems Engineering*, 153, 110–128.
- Becker-Reshef, I. (2010). Monitoring global croplands with coarse resolution earth observations: The Global Agriculture Monitoring (GLAM) project. *Remote Sensing*, 2(6), 1589–1609, 22.
- Blackmore, S., Stout, B. A., Wang, M., & Runov, B. (2005). Robotic agriculture—The future of agricultural mechanization?. In: *European conference on precision agriculture* (Vol. 5; pp. 621–628). Uppsala, Sweden: Wageningen Academic Publishers.
- Buning, E. A. (2010). Electric drives in agricultural machinery—approach from the tractor side. *Journal of Agricultural Engineering*, 47(3), 30–35.
- Cbinsights. (2019). *AgTech deal activity more than triples*. Available from <https://www.cbinsights.com/research/agriculture-farm-tech-startup-funding-trends/>. Accessed 18.02.19.
- Chedad, A. (2001). AP—Animal production technology: Recognition system for pig cough based on probabilistic neural networks. *Journal of Agricultural Engineering Research*, 79(4), 449–457.
- Conesa-Muñoz, J., Gonzalez-de-Soto, M., Gonzalez-de-Santos, P., & Ribeiro, A. (2015). Distributed multi-level supervision to effectively monitor the operations of a fleet of autonomous vehicles in agricultural tasks. *Sensors*, 15, 5402–5428.
- Coopersmith, E. J., Minsker, B. S., Wenzel, C. E., & Gilmore, B. J. (2014). Machine learning assessments of soil drying for agricultural planning. *Computers and Electronics in Agriculture*, 104, 93–104.
- Corke, P. I., Roberts, J. M., & Winstanley, G. J. (1998). Robotics for the mining industry. In A. T. de Almeida, & O. Khatib (Eds.), *Autonomous robotic systems. Lecture notes in control and information sciences* (236). London: Springer.
- Craninx, M., Fievez, V., Vlaeminck, B., & De Baets, B. (2008). Artificial neural network models of the rumen fermentation pattern in dairy cattle. *Computers and Electronics in Agriculture*, 60, 226–238.
- Daroya, R., & Ramos, M. (2017). NDVI image extraction of an agricultural land using an autonomous quadcopter with a filter-modified camera. In: *Proceedings of the 2017 7th IEEE international conference on control system, computing and engineering (ICCSCCE)* (pp. 110–114). Penang, Malaysia.
- Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., . . . Henry, D. (2015). Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture*, 111, 18–28.
- Esmeyjer, J., Bakker, T., Ooms, M., & Kotterink, B. (2015). *Data-driven innovation in agriculture: Case study for the OECD KBC2-programme*. TNO report TNO 2015 R10154.
- Griepentrog, H. W., Nrremark, M., Nielsen, H., & Blackmore, B. S. (2005). Seed mapping of sugar beet. *Precision Agriculture*, 6, 157–165.
- Gutiérrez, P. (2008). Logistic regression product-unit neural networks for mapping *Ridolfia segetum* infestations in sunflower crop using multitemporal remote sensed data. *Computers and Electronics in Agriculture*, 64(2), 293–306.
- Hansen, M. F., Smith, M. L., Smith, L. N., Salter, M. G., Baxter, E. M., Farish, M., . . . Grieve, B. (2018). Towards on-farm pig face recognition using convolutional neural networks. *Computers & Industrial Engineering*, 98, 145–152.
- Hayashi, S., Yamamoto, S., Shigematsu, K., Kobayashi, K., Kohno, Y., Kamata, J., . . . Kurita, M. (2011). Performance of movable-type harvesting robot for strawberries. *Acta Horticulturae*, 893, 317–324. Available from <https://doi.org/10.17660/ActaHortic.2011.893.27>.
- Henten, E.J., Van, Bac, C.W., Hemming J., & Edan Y. (2013). Robotics in protected cultivation. In: *IFAC proceedings volumes* <https://doi.org/10.3182/20130828-2-SF-3019.00070>.
- Johann, A. L., de Araújo, A. G., Delalibera, H. C., & Hirakawa, A. R. (2016). Soil moisture modeling based on stochastic behavior of forces on a no-till chisel opener. *Computers and Electronics in Agriculture*, 121, 420–428.
- Kamilaris, A., Kartakoullis, A., & Prenafeta Boldú, F. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, 143. Available from <https://doi.org/10.1016/j.compag.2017.09.037>.
- Kempenaar, C., Lokhorst, C., Bleumer, E. J. B., & Veerkamp, R. F. (2016). *Big data analysis for smart farming*, . (655). Wageningen University & Research.
- Khanal, S., Fulton, J., & Shearer, S. (2017). An overview of current and potential applications of thermal remote sensing in precision agriculture. *Computers and Electronics in Agriculture*, 139, 22–32.
- Lamborelle, A., & Fernández Álvarez, L. (2019). *Farming 4.0: The future of agriculture?*
- Lucas, M. T., & Chhajed, D. (2004). Applications of location analysis in agriculture: A survey. *Journal of the Operational Research Society*, 55(6), 561–578.
- Luck, J. D., & Fulton, J. P. (2015). Improving yield map quality by reducing errors through yield data file post-processing. *Institute of Agriculture and Natural Resources*, 9.

- Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., . . . Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, *51*, 47–60.
- Madhuri, J., & Indiramma, M. (2019). Role of big data in agriculture. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(2), 12–21.
- Mahajan, U., & Raj, B. (2016). Drones for Normalized Difference Vegetation Index (NDVI), to estimate crop health for precision agriculture: A cheaper alternative for spatial satellite sensors. In: *Proceedings of the international conference on innovative research in agriculture, food science, forestry, horticulture, aquaculture, animal sciences, biodiversity, ecological sciences and climate change* (AFHABEC-2016). Delhi, India, 22 October 2016.
- Manyica, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2019). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey.
- McBratney, A. B., & Whelan, B. M. (1999). *The null hypothesis of precision agriculture*. Sheffield, UK: Sheffield Academic Press.
- Mohammadi, K., Shamshirband, S., Motamedi, S., Petković, D., Hashim, R., & Gocic, M. (2015). Extreme learning machine based prediction of daily dew point temperature. *Computers and Electronics in Agriculture*, *117*, 214–225.
- Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotziou, G., . . . Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, 104–116.
- Mucherino, A., Papajorgij, P., & Pardalos, P. (2009). Data mining in agriculture. *Springer Science & Business Media*, *34*.
- Murakami, Y., Utomo, S., Hosono, K., Umezawa, T., & Osawa, N. (2013). iFarm: Development of cloud-based system of cultivation management for precision agriculture. In: *2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE)* (pp. 233–234).
- Murugesan, R., Sudarsanam, S. K., Malathi, G., Vijayakumar, V., Neelanarayanan, V., Venugopal, R., . . . Malolan, V. (2019). Comparison of corruption prevention system around the world 2. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(2), 1870–1877.
- Nidhi. (2020). *Big data for smart agriculture, . Smart village technology, modeling and optimization in science and technologies* (17, pp. 181–189). Springer, Cham.
- Patel, R., & Patel, M. (2013). Application of cloud computing in agricultural development of rural India. *International Journal of Computer Science and Information Technologies*, *4*(6), 922–926.
- Patil, A. P., & Deka, P. C. (2016). An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Computers and Electronics in Agriculture*, *121*, 385–392.
- Pedersen, S. M., Fountas, S., & Blackmore, S. (2008). *Agricultural robots-applications and economic perspectives: Chapter 21. Service robot applications* (pp. 369–382). UK: IntechOpen.
- Pegorini, V., Karam, L. Z., Pitta, C. S. R., Cardoso, R., da Silva, J. C. C., Kalinowski, H. J., . . . Assmann, T. S. (2015). In vivo pattern classification of ingestive behavior in ruminants using FBG sensors and machine learning. *Sensors*, *15*, 56–71.
- Reddy, N., Reddy, A., & Kumar, J. A. (2016). Critical review on agricultural robots. *International Journal of Mechanical Engineering and Technology (IJMET)*, *7*, 6.
- Reeve, D. R., Eizad, Z., & Ramm, A. F. (2011). *Method for decomposing task e.g. crop spraying task, to be performed on e.g. agricultural field by e.g. tractor-puller sprayer vehicle assembly, involves decomposing top-order layer based on rules to form bottom-order layer. US2011257850-A1*.
- Rudd, J.D., Roberson, G.T., & Classen, J.J. (2017). Application of satellite, unmanned aircraft system, and ground-based sensor data for precision agriculture: A review. In *Proceedings of the 2017 ASABE Annual International Meeting. American Society of Agricultural and Biological Engineers*. Spokane, WA, USA.
- Saiz-Rubio, V., & Rovira-Más, F. (2020). From smart farming towards Agriculture 5.0: A review on crop data management. *Agronomy*, *10*(2), 207. Available from <https://doi.org/10.3390/agronomy10020207>.
- Scott, J. (2010). Aerial seeding of cover crops. U.S. Department of Agriculture, Natural Resources Conservation Service (NRCS). Iowa State Office, Des Moines, Iowa. September 2010.
- Shafi, U., Mumtaz, R., García-Nieto, J., Hassan, S. A., Zaidi, S. A. R., & Iqbal, N. (2019). Precision agriculture techniques and practices: From considerations to applications. *Sensor*, *19*, 2–25.
- Shamshiri, R. R., Weltzien, C., Hameed, I. A., Yule, I. J., Grift, T. E., Balasundram, S. K., . . . Chowdhary, G. (2018). Research and development in agricultural robotics: A perspective of digital farming. *International Journal of Agricultural and Biological Engineering*, *11*, 1–14.
- Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., & Skakun, S. (2017). Exploring Google Earth engine platform for big data processing: Classification of multi-temporal satellite imagery for crop mapping. *Frontiers of earth science*, *5*, 17.
- Skakun, S., Justice, C. O., Vermote, E., & Roger, J. C. (2018). Transitioning from MODIS to VIIRS: An analysis of inter-consistency of NDVI data sets for agricultural monitoring. *International Journal of Remote Sensing*, *39*, 971–992.
- Sonka, S. (2014). Big data and the Ag sector: More than lots of numbers. *International Food and Agribusiness Management Review*, *17*, 1–20.
- Varadharajan, D. AI. (2019). *Robotics, and the future of precision agriculture*.
- Verified Market Intelligence. (2018). *Global agriculture robots. Market size, status and forecast to 2025*. Boonton, NJ, USA: Verified Market Intelligence Inc.
- Veroustraete, F. (2015). The rise of the drones in agriculture. *EC Agriculture*, *2*, 325–327. Available from <https://doi.org/10.1016/j.jocn.2013.06.004>.
- Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J. A., & Buytaert, W. (2015). Web technologies for environmental big data. *Environmental Modelling & Software*, *63*, 185–198.
- Walch, K. (2020). *How AI is transforming agriculture*.
- Wheeler, T., & von Braun, J. (2013). Climate change impacts on global food security. *Science (New York, N.Y.)*, *341*(80), 508–513. Available from <https://doi.org/10.1126/science.1239402>.

- Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, 2017, 1–17. Available from <https://doi.org/10.1155/2017/1353691>.
- Yamamoto, S., Hayashi, S., Saito, S., Ochiai, Y., Yamashita, T., & Sugano, S. (2010). Development of robotic strawberry harvester to approach target fruit from hanging bench side. *IFAC Proceedings*, 43, 95–100. Available from <https://doi.org/10.3182/20101206-3-JP-3009.00016>.
- Yan-e, D. (2011). Design of intelligent agriculture management information system based on IoT, In: *fourth international conference on intelligent computation technology and automation* (pp. 1045–1049). Shenzhen, Guangdong.
- Zhang, Q. (2015). *Precision agriculture technology for crop farming* (1st ed.). Boca Raton, FL, USA: CRC Press and Taylor & Francis Group. ISBN 978-1-4822-5107-4.
- Zhang, Y. (2019). The role of precision agriculture. *Resource*, 19, 9.
- Zhao J., Zhang J.F., Feng Y. & J Guo (2010). The study and application of the IOT technology in agriculture. In: *2010 3rd international conference on computer science and information technology* (pp. 462–465). Chengdu.

This page intentionally left blank

Artificial intelligence: a way forward for agricultural sciences

Neeru S. Redhu^{1,*}, Zoozeal Thakur^{2,*}, Shikha Yashveer¹ and Poonam Mor³

¹Department of Molecular Biology, Biotechnology & Bioinformatics, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India, ²Bacteriology Lab, National Research Centre on Equines, Hisar, Haryana, India, ³Department of Languages & Haryanvi Culture, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India

37.1 Introduction of artificial intelligence

Artificial intelligence (AI) is a combination of conventional science disciplines, scientific theories, and practices using mathematical logic, statistics, probabilities, through computers to imitate the cognitive abilities of humans. In other words, AI is a subfield of computer science which deals with the creation of tangible or intangible systems which not only behave intelligently but also display behavior similar to human beings including speech recognition, natural language understanding and translation, knowledge management, image analysis, decision making, and learning among others. Achieving human-like performance in all cognitive tasks using purely logical reasoning makes systems powerful and useful. Thus the “artificial” in AI can be understood as “nonbiological,” the “intelligence” can be taken as “ability to accomplish complex goals or tasks” (Brewka, 1996).

The term AI was designed by John McCarthy in 1955 and defined by Marvin Minsky as *the construction of computer programs that engage in tasks that are currently more satisfactorily performed by human beings because they require high-level mental processes such as perceptual learning, memory organization, and critical reasoning*. The Rockefeller Institute-funded conference in 1956 at Dartmouth College was the first step toward its foundation (Brewka, 1996). The advancement in AI is closely linked with the developments in the computing field. Started with simple “if-then rules,” the AI has currently progressed to behaving like a human brain using a variety of complex algorithms. Technology has advanced at such an accelerating pace that we have computers in our pockets that are connected to the Internet which provides plenty of information at our fingertips along with a plethora of other options like streaming video and music at any moment. AI involves both the collection and organization of large amounts of data to attain insights and to make predictions using above human capabilities (Sajja, 2021).

Fundamentally, AI aims to achieve thinking like humans, acting like a human, thinking rationally, and acting rationally; therefore it has been driven by all of the four objectives by employing and developing different methods. A human-centric aim is generally fulfilled by an empirical approach involving human behavioral observation and hypotheses. While the rational-centric aim is achieved by a combination of mathematics and engineering. Thinking like a human is mainly following a cognitive modeling approach; combining both AI models and experimental psychology techniques to hypothesize precise human mind/thinking process (Brewka, 1996). Acting like a human is mainly described by Turing and the Total Turing Test (TT) (see Box 37.1). For achieving this, the computer must have natural language processing (NLP), automation reasoning, machine learning (ML), knowledge representation, computer vision, and robotics capabilities. Thinking rationally involves the use of logic. For instance, “XYZ is a man; all men are mortal; therefore XYZ is mortal.” Logistic programs developed in 1965 were abiding by this approach and in principle to any solvable problem. Acting rationally primarily involves achieving the best outcome or correct inferences (Sajja, 2021).

From Apple’s intelligent personal assistant Siri or Cortana of Microsoft to self-driving cars, AI is improving swiftly. Other popular examples include Google Maps, ridesharing in cabs, Face recognition in Facebook photo upload, Face

*Equal authorship.

BOX 37.1 Turing Test

The “Turing Test” (TT) involves a human and a computer in two sealed rooms, and a human judge to determine in which of the two rooms contains a human and a computer by asking questions by email (originally, it was by teletype messaging). If, after receiving answers, the judge can not perform better than 50/50 in recognizing the room of the human and the computer, it can be said that the computer has passed the TT. Passing the test in this method operationalizes linguistic indistinguishability. Later, Turing explicitly suggested that the “child machines” be built and that these machines could then gradually grow up on their own to learn to communicate in natural language at the level of adult humans. Turing test deliberately avoided direct physical interaction between human judge and computer. As the physical simulation is considered unnecessary for intelligence. But the “Total Turing Test” involves a video signal; so that judge can examine the perpetual abilities and have an opportunity to judge based on complete interactions.

unlock in Mobile, search, and recommendation in online shopping sites. Currently, Google with the goal of “AI-first” world is using AI technologies for numerous applications and operating two of the top AI research labs in the world, that is, DeepMind in London and GoogleBrain in California (Davenport & Ronanki, 2018). Robotics, one of the major AI field, require intelligence for operating jobs involving object manipulation, motion planning, mapping, and navigation. An added advantage, the AI systems, once programmed to perform and evolve for doing specific tasks are unbiased and this could have a positive impact on the interaction between AI systems and the society at large. The AI applications are not limited to any discipline and can also help in agriculture by detecting diseases, reducing agricultural risks, predicting consumer behavior, and helping farmers increase crop yields among others (Murase, 2000).

AI can be utilized to making intelligent embedded systems that are responsive like humans and can work quickly with higher precision. AI, jointly with automation, Internet of Things (IoT) devices, and solar-powered and sensor technology facilitates precision and climate-smart agriculture. Besides, AI techniques such as expert and mobile-based recommender systems can also significantly enhance the adoption of AI in agriculture particularly for high-yielding or disease-resistant varieties and innovative farm technique implementation (Zhao, 2020). AI can also help farmers to maximize their cultivable field, by providing precise information about the types of crops, weather patterns, and best conditions for crop cultivation. AI techniques like machine and deep learning (DL) are being used effectively on image data for segmentation for disease/variety identification, crop yield, field monitoring, and predicting the time of application and optimum dose of chemical sprays, time of harvest, and life of produce among others (Murase, 2000; Pantazi et al., 2020). AI research, tools, and technology are evolving every day and reaching new horizons. In the last five years, the reported annual AI growth is 12.9% across the world, which is truly commendable.

37.2 History of artificial intelligence

The notion of simulating intelligent behavior and critical thinking by computers was first expressed by Alan Turing in 1950 in the book entitled “Computers and Intelligence,” as a test that determines whether computers have abilities to achieve intelligence similar to humans or not. Even though at that time it was not called AI; Still, John Von Neumann and Alan Turing are considered as the founding fathers of AI technology (Haenlein & Kaplan, 2019). They standardized the architecture of our contemporary computers and made the transition from machines to binary logic and computers to decimal logic. They demonstrated that computer capabilities for executing whatever are they are instructed or programmed to do. Turing, then further rose the question “why machine can’t use available information as well as the reason to solve problems and make decisions like a human?” He further discussed building intelligent machines and testing their intelligence in an article entitled *Computing Machinery and Intelligence* in 1950 (Turing, 2012). Turing described it as a “game of imitation,” that involves a human differentiating between a man or a machine via teletype chatting and designed a TT (see Box 37.1). The article, even though called controversial at times, is often cited as the beginning of AI and the questioning of the human and the machine boundary lines (Turing, 2012).

The developments of AI coincide with technological progress and the desire to achieve the functioning of machines to the human levels. The first mathematical and computer model of the neurons developed in 1943 by Warren McCulloch and Walter Pitts along with the unification of mathematical theory, electronics, and automation via cybernetics by Norbert Wiener in 1948 marked the initial effort for AI (Muthukrishnan et al., 2020). The computer at that time lacks a prerequisite for intelligence, that is, memory, hence could only execute commands but couldn’t remember them; therefore cannot develop any further understanding. The Logic Theorist program of 1956 created by Allen Newell, Cliff Shaw, and Herbert Simon emulates human problem-solving skills. This is deemed as the first AI program and was presented in 1956 at the

Dartmouth Summer Research Project on Artificial Intelligence conference hosted by John McCarthy and Marvin Minsky (Muthukrishnan et al., 2020; Wilamowski & Irwin, 2016). Even though this historic AI conference brought top researchers from various fields together and had an open-ended discussion, it fell short of achieving anything as there was no agreement on the standard methods for AI advancement (Muthukrishnan et al., 2020; Wilamowski & Irwin, 2016). Nonetheless, everybody agreed with the AI achievability and catalyzed the next twenty years of AI research.

Even though AI was fascinating and promising, its popularity declined in the early 1960s. Owing to memory constraints in the early machines, it was difficult to attain any of the AI goals. However, the foundations such as information processing language (the basis of logic theorist machine program) and solution trees were laid (Millstein, 1968). AI flourished with the advent of the first microprocessors as computers now are cheaper, faster, and accessible with higher storage. ML algorithms based “inference engine” was developed to mirror human logical reasoning and to help people in choosing algorithms according to their problem. For instance, DENDRAL of MIT in 1965 and MYCIN of Stanford University in 1972 are developed; these are the specialized system for molecular chemistry and diagnosis of blood diseases as well as prescription drugs respectively. Despite everything, the lacuna before achieving the end goal of AI, that is, machine with the general average intelligence of a human is still present (Haenlein & Kaplan, 2019; Muthukrishnan et al., 2020).

During the 1980s, the development of the algorithmic toolkit using DL and expert systems as well as increased funding, again reignited the AI. DL techniques developed by John Hopfield and David Rumelhart allow a computer to learn from experience whereas Edward Feigenbaum’s expert systems simulated the decision-making process of a human expert (Haenlein & Kaplan, 2019; Muthukrishnan et al., 2020). From 1982–90, the Japanese government invested \$400 million in expert systems and other AI-related endeavors under Fifth Generation Computer Project to revolutionize computer processing, logic program implementation, and improving AI (Jaakkola et al., 2019). Unfortunately, the majority of the objectives were not achieved and funding ceased, resulting in reducing in AI popularity. The main problem was the absence of understanding about machine reasoning that in turn caused difficulty in AI development. Additionally, faster, cheaper, and simpler methods were developed to solve problems. This gives rise to the term advanced computing in the 1990s (Haenlein & Kaplan, 2019; Muthukrishnan et al., 2020).

Ironically, in the lack of public scrutiny, AI bloomed and numerous goals had been accomplished such as intelligent decision-making, speech recognition, and kismet robot. In 1997, the winning of IBM’s chess-playing program named Deep Blue against the world chess champion and grandmaster Gary Kasparov served as a great leap toward AI development (Haenlein & Kaplan, 2019; Muthukrishnan et al., 2020). The program utilizes a systematic brute force algorithm where all possible moves were scored. Similarly, Dragon Systems’ speech recognition software was another huge step in AI’s endeavor in spoken language interpretation. The Kismet robot by Cynthia Breazeal could recognize and display emotions is another success of AI (Breazeal, 2003). All this could be achieved due to the development in computer storage and processing speed, key limiting factors in early AI research. Still, each program is only able to manage edge in a specific field with few parameters as input which do not represent the full scale of complexity in the world (Haenlein & Kaplan, 2019; Muthukrishnan et al., 2020).

The development of AI is closely linked with the invention of the computer system. As the advent of computers increased the fundamental limit of storage the AI achieved many of its goals. Even though the scale is limited still it is a huge step in the forward direction. With the further advent of DL in the 2000s and other ML techniques, AI has become capable of analyzing complex algorithms, and decision making. Google’s AlphaGo who defeated the Chinese Go champion, Ke Jie in 2016 is another example of an AI success story (Silver et al., 2017). The game “Go” is the most challenging classical game for AI due to its complexity. Alpha Go used deep neural networks and reinforcement learning for decision making. New bloom in the discipline since 2010 can be attributed to the access to massive volumes of data as well as to the invention of graphics card processor that enhances the efficiency of learning algorithms (Chen, 2016b). Table 37.1 illustrates major milestones in AI development.

37.3 Methods and approaches in artificial intelligence

Today, AI has grown to be a significant component of the technology industry and its highly specialized and technical research. The core of AI includes computer programs targeting problems involving reasoning, knowledge, problem solving, learning, ability to manipulate and move objects among others.

Currently, AI can be classified into three types: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Super Intelligence (ASI) (Chang et al., 2018). ANI refers to the ability of a computer in performing a single task extremely well, such as playing chess or finding efficient routes to places while riding a car by Google Maps. AGI comes to play in performing any intellectual task like reasoning, solving problems, and making judgments along with planning, learning, and integrating prior information in decision-making as well as innovative, and imaginative

TABLE 37.1 Major milestones in artificial intelligence history.

Year	Significance in artificial intelligence
1763	Thomas Bayes develops a framework for reasoning about the probability of events. The Bayesian inference will become a leading approach in machine learning.
1914	The Spanish engineer Leonardo Torres y Quevedo demonstrates the first chess-playing machine, capable of king and rook against king endgames without any human intervention.
1921	Czech writer Karel Čapek introduces the word “robot” in his play R.U.R. (Rossum’s Universal Robots). The word “robot” comes from the word “robota” (work).
1943	Warren S. McCulloch and Walter Pitts published “A Logical Calculus of the Ideas Immanent in Nervous Activity” in the Bulletin of Mathematical Biophysics. They discussed networks of idealized and simplified artificial “neurons” and how they might perform simple logical functions. (This will become the inspiration for computer-based “neural networks” and later “deep learning.”)
1950	Alan Turing publishes “Computing Machinery and Intelligence” in which he proposes “the imitation game” which will later become known as the “Turing Test.”
1955	The term “artificial intelligence” is coined in a proposal submitted by John McCarthy (Dartmouth College), Marvin Minsky (Harvard University), Nathaniel Rochester (IBM), and Claude Shannon (Bell Telephone Laboratories). The workshop, in July and August 1956, is generally considered as the official birthdate of AI.
1957	Frank Rosenblatt develops the Perceptron, an early artificial neural network enabling pattern recognition based on a two-layer computer learning network.
1958	John McCarthy develops the programming language Lisp which becomes the most popular programming language used in artificial intelligence research.
1965	Joseph Weizenbaum develops ELIZA, an interactive program that carries on a dialog in the English language on any topic. Weizenbaum, who wanted to demonstrate the superficiality of communication between man and machine, was surprised by the number of people who attributed human-like feelings to the computer program.
	Edward Feigenbaum, Bruce G. Buchanan, Joshua Lederberg, and Carl Djerassi start working on DENDRAL at Stanford University. The first expert system, automated the decision-making process and problem-solving behavior of organic chemists.
1969	Arthur Bryson and Yu-Chi Ho describe backpropagation as a multistage dynamic system optimization method. A learning algorithm for multilayer artificial neural networks has contributed significantly to the success of deep learning, once computing power has sufficiently advanced to accommodate the training of large networks.
1970	The first anthropomorphic robot, the WABOT-1, is built at Waseda University in Japan. It consisted of a limb-control system, a vision system, and a conversation system.
1972	MYCIN, an early expert system for identifying bacteria causing severe infections and recommending antibiotics, is developed at Stanford University
1978	The XCON (eXpert CONfigurer) program, a rule-based expert system assisting in the ordering of DEC’s VAX computers by automatically selecting the components based on the customer’s requirements, is developed at Carnegie Mellon University.
1986	The first driverless car, a Mercedes-Benz van equipped with cameras and sensors, built at Bundeswehr University in Munich by Ernst Dickmanns, drives up to 55 mph on empty streets.
	David Rumelhart, Geoffrey Hinton, and Ronald Williams describe “a new learning procedure, back-propagation, for networks of neuron-like units.”
1988	Judea Pearl publishes Probabilistic Reasoning in Intelligent Systems. He was awarded the 2011 Turing Award.
	Rollo Carpenter develops the chat-bot Jabberwacky to “simulate natural human chat in an interesting, entertaining and humorous manner.” It is an early attempt at creating artificial intelligence through human interaction.
	Members of the IBM T.J. Watson Research Center heralded the shift from rule-based to probabilistic methods of machine translation and reflecting a broader shift to “machine learning” based on statistical analysis.
1989	Yann LeCun and other researchers at AT&T Bell Labs successfully apply a backpropagation algorithm to a multilayer neural network, recognizing handwritten ZIP codes. Given the hardware limitations at the time, it took about 3 days to train the network.
1995	Richard Wallace develops the chatbot A.L.I.C.E (Artificial Linguistic Internet Computer Entity), inspired by Joseph Weizenbaum’s ELIZA program, but with the addition of natural language sample data collection on an unprecedented scale.

(Continued)

TABLE 37.1 (Continued)

Year	Significance in artificial intelligence
1997	Sepp Hochreiter and Jürgen Schmidhuber propose Long Short-Term Memory (LSTM), a type of recurrent neural network used today in handwriting recognition and speech recognition.
	Deep Blue becomes the first computer chess-playing program to beat a reigning world chess champion.
1998	Dave Hampton and Caleb Chung create Furby, the first domestic or pet robot.
2000	MIT's Cynthia Breazeal develops Kismet, a robot that could recognize and simulate emotions.
	Honda's ASIMO robot, an artificially intelligent humanoid robot, can walk as fast as a human, delivering trays to customers in a restaurant setting.
2001	A.I. Artificial Intelligence is released, a Steven Spielberg film about David, a childlike android uniquely programmed with the ability to love.
2007	Fei Fei Li and colleagues at Princeton University start to assemble ImageNet, a large database of annotated images designed to aid in visual object recognition software research.
2009	Computer scientists at the Intelligent Information Laboratory at Northwestern University develop Stats Monkey, a program that writes sport news stories without human intervention.
2011	A convolutional neural network wins the German Traffic Sign Recognition competition with 99.46% accuracy (vs humans at 99.22%).
2014	The first driverless car designed by Google to pass a self-driving test in Nevada, United States.
2016	Google DeepMind's AlphaGo defeats Go champion Lee Sedol.
2018	Microsoft's Project Brainwave based on deep learning for real-time AI inference in the cloud and on the edge was launched.

similar to humans by a computer program. As AI is getting powerful day by day; ASI is when a computer or a system surpasses human intellect, that is, it is wiser, creative, more socially adept, and better as well as smarter than the sum of all humanity combined (Sundvall, 2019). ML, Robotics, NLP, Automated Reasoning, Expert Systems, Computer Vision, Speech Recognition, Automated Data Analytics, Virtual Reality, Augmented Reality, IoT, Cloud Computing, DL among others are some major subareas of AI having huge potential in solving complex problems of agriculture (Bundy, 2017). Fig. 37.1 depicts the relationship between AI methods, approaches, algorithms, and subfields.

Currently, there are numerous methods popular for AI-driven technologies and systems. The methods of AI include ML, DL, and artificial neural network (ANN) among others.

37.3.1 Machine learning

ML is the subfield of AI that uses previously obtained data to recognize patterns that can be used for further analysis of specific data. The machine, therefore “learns” and applies that information dynamically to future similar scenarios. In other words, it enables automatic learning and improvement by the system without being explicitly programmed and without human intervention or assistance (Liakos et al., 2018). With the aid of ML practical speech recognition, self-driving cars, effective web search, and insight into the genome have been achieved. Nowadays, ML is so prevalent that one probably uses it throughout the day without even realizing it. Using algorithms and neural network models ML constructs a mathematical model based on sample data (training data) to assist in progressively improving the performance of computer systems to make predictions or decisions (Tu, 2019).

The idea behind ML is based on a brain-cell interaction model developed using theories of communication between neurons. It was first described by Donald Hebb in a book entitled *The Organization of Behavior* in 1949 (Shaw, 1986). In the 1950s, Arthur Samuel of IBM created a computer program for playing checkers which had a scoring function associated with the positions of the board pieces. The program decides its next step following a minimax strategy (later becomes minimax algorithm). He further modified the program in such a way that it can become better by recording/remembers already seen positions and scoring function (Devroye & Lugosi, 2001). The term “Machine Learning” was given by Arthur Samuel in 1952 (Panesar & Panesar, 2019). Since then, ML algorithms and methodologies were

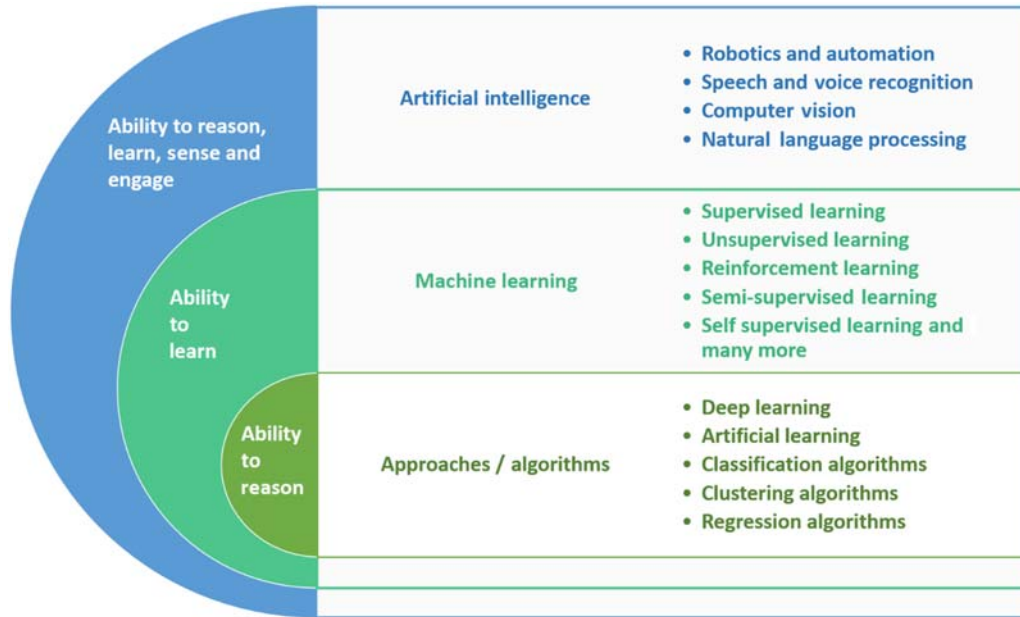


FIGURE 37.1 Relationship between artificial intelligence (AI) methods, approaches, algorithms, and subfields. AI, machine learning, and algorithms have a very interdependent relationship and often time might be confusing. Algorithms are a specific set of rules for a particular task, approaches might include more than algorithms. Machine learning can use one or more approaches for achieving the desired target. AI can be achieved using machine learning or artificial neural network or deep learning. Robotics, automation, computer vision, natural language processing, speech and voice recognition are the application of AI that are beneficial for a variety of purposes.

developed in the form of neural network (1957), nearest neighbor algorithm (1967), feedforward/multilayered neural networks (the 1960s), backpropagation (1970s), ANN (1980s), and DL among others (Zhou, 2015). Until the early 1980s, ML was used as a training program for AI; but later AI research started focusing on using only logical and knowledge-based approaches. This created a division between AI and ML. Still, ML researchers continued to work in the field and with the advent of boosting algorithms developed by Robert Schapire in 1990, ML flourished (Jordan & Mitchell, 2015). Boosting algorithms reduces the bias during learning and transforms a set of weak points/learners (i.e., classifiers which are vaguely correlated with their true classification) into a single strong point/learner (i.e., classifiers which are properly aligned with their true classification) through repetitive learning (Schapire, 2013).

Model creation is the significant feature of performing ML wherein input data is trained so that they can process additional data to make decisions. Modern ML models are adept in continuously learning, thus becoming more precise the longer they run and with new computing technologies, they also have higher scalability and efficiency. From 1990 to date speech recognition, facial recognition and self-driving vehicles are some of the success stories of ML (Jordan & Mitchell, 2015). The newer concepts and technologies derived from the ML include new algorithms for robots, IoT, analytics tools, chatbots, and more. Currently, ML models are used for a variety of predictions, that is, from disease outbreaks to the rise and fall of stocks to fraud detection to product recommendations to customer personalization, to data analysis of streamlined data, and many more (Mohammed et al., 2016). Besides, the ML models are also being developed and employed to forecast environmental impacts on crop yield due to weather changes as well as crop sustainability prediction and detection of potential diseases and pests.

ML algorithms are often categorized based on the input data, output data, and the problem they are proposed to solve. Three methods of ML are supervised, unsupervised, and reinforcement learning (Vieira et al., 2019). Other methods of ML such as semisupervised, self-supervised, multiinstance, inductive, transductive learning among others are mostly variants of these three.

37.3.1.1 Supervised learning

Supervised learning includes both input variables (X) and an output variable (Y) and a mapping function ($Y = f(X)$) which links the input to the output during learning through an algorithm. The aim is to accurately estimate the mapping function in a way that the prediction of output variables (Y) is also precise when the new input data (X) was introduced (Cunningham et al., 2008). Supervised learning algorithms construct a mathematical model of a dataset (training data)

comprising inputs, desired outputs/outcomes as well as training examples (i.e., previously designed models). The training example and training data in the mathematical model are depicted by the array (also called feature vector) and matrix respectively. The name “Supervised learning” reflects the nature of the learning process, that is, we know the outcome and the algorithm has to make predictions repetitively till it achieves the desired outcome. The repeated prediction of the mapping function causes its optimization, training and finds the optimal function that precisely determines the outcomes for inputs that were not present in the training data. Improved accuracy of outcomes by an algorithm over time is recognized as learning for the execution of the task. The majority of practical ML models are created using supervised learning (Kotsiantis, 2007).

Supervised learning algorithms are categorized into classification and regression. Classification algorithms are employed when the outputs possibly fall into a limited set of predefined categories; whilst, regression algorithms are employed when the outcome possibly has a numerical value within a range. For training of a classification algorithm, the data points and an assigned category or class are provided (Radhakrishnan et al., 2007). Following which the classification algorithms assign a class/category to an input value, according to the training data provided. For example, for determining whether an email is spam or not; spam and not spam are considered as two classes and a classification algorithm will be provided emails belonging to both of these classes (training data). The supervised learning algorithm model then identifies data features correlated to either class and created a mapping function ($Y=f(X)$). The mapping function of the model, on encountering a new email, establishes whether the new email is spam or not. Classification algorithms include linear classifiers, decision trees, support vector machines (SVM), k-nearest neighbor, and random forest; the choice of algorithms depends on the data (Kotsiantis, 2007; Vieira et al., 2019).

Unlike classification, regression establishes the significant relationship between dependent and independent variables (Radhakrishnan et al., 2007). For example, for determining a student’s test grade according to the number of hours per week spent in studying; here hours per week is an independent variable while test score is a dependent variable. The regression algorithm will determine the correlation between them; so that a line for best fit can be plotted using the data points representing model predictions which will be used to predict the test score of new student’s. The regression algorithms include linear regression, polynomial regression, and logistic regression (Kotsiantis, 2007).

37.3.1.2 Unsupervised learning

Contrary to supervised learning, unsupervised learning uses only input data (X) but no corresponding output variables for developing mapping function. Hence training takes place without labeled outcome or supervision. The unsupervised algorithms mainly assemble unsorted data according to similarities, and differences and identify hidden patterns and structures in data by themselves. It is known as unsupervised learning due to the absence of predefined outcomes of supervision (Francis, 2014; Zheng, 2015). They recognize common features in the input data and respond based on the presence or absence of these common features in new input data. For example, if an image contains fox and lion, then during learning the algorithm or machine can not categorize it as fox or lion; as it doesn’t have any idea or information about them. Still, it can classify them as per their similarities, patterns, and differences. So, with the aid of unsupervised learning, the computer model discovers previously undetected patterns and information on its own.

The unique feature of unsupervised learning algorithms is that it learns from test data without being labeled, classified or categorized. Unsupervised learning algorithms can be classified into two, that is, clustering and association categories. The clustering or cluster analysis is performed when the desired outcome is to divide data into various groups such as customers according to their purchasing behavior or viewers according to their watched list. Cluster analysis is the most common method of unsupervised learning which is primarily used for exploratory data analysis. The clustering algorithms include k-means clustering, self-organizing maps, Gaussian mixture models, hierarchical clustering, and hidden Markov models among others (Xu & Tian, 2015). While, the association is performed to establish rules that correlate large portions of input data like buyers of X also have a tendency to buy Y or viewer of X also have a tendency to watch Y. Other unsupervised algorithms are k-nearest neighbors, hierarchal clustering, anomaly detection, a priori algorithm and many more (Moutinho et al., 2014). Unsupervised learning methods are prominently used in bioinformatics studies like sequence analysis, sequence data mining, pattern mining, and genetic clustering; in medical imaging and computer vision for image segmentation and object recognition respectively (Francis, 2014; Topol, 2019).

37.3.1.3 Reinforcement learning

Reinforcement learning is based on the human brain’s “trial and error” or “learning from their mistake” learning mechanism. It uses computational power and software for developing a model in an interactive environment by utilizing feedback from its actions and experiences (Bhatnagar et al., 2013). In simpler words, the reinforcement learning method

follows a series of decision making whilst every step was taken by the model is awarded reward points and the model will accumulate all the reward points based on steps taken before achieving the end goal. Similar to the video games where players gather scores that will increase their level one at a time; the objective of the reinforcement learning algorithm is to identify the next correct answer that will take it to the next step of the process (Szepesvári, 2010).

Similar to supervised learning, it also derives mapping function using input and output variables, but contrary to supervised learning it uses awards and penalties for positive and negative performance respectively. In terms of comparison with unsupervised learning which focuses on finding similarities and differences amongst training data, reinforcement learning has an entirely different objective, that is, to find an appropriate model with maximum cumulative award points (Tu, 2019). The basic elements of the reinforcement learning model are environment, state, reward, policy, and value; referring to the physical world for the agent operation, the current situation of the agent, feedback from the environment, the method to map agent's state to actions and possible future rewards basis on action taken in a particular state respectively (Bhatnagar et al., 2013). For instance, in the PacMan game, the PacMan (agent) eats the dots in the grid while evading the ghosts. Here, the grid reflects the interactive environment; scores received on eating dots are rewards, and loss of game and life is the penalty for getting killed by the ghost. The states are the PacMan location within the grid and the total cumulative reward is a game win. The policy makes PacMan explore new states and reward maximizing along with trying to find the optimal policy in turn (Tu, 2019; Vieira et al., 2019).

The reinforcement learning approach can be categorized into model-based and model-free approaches. Model-based approaches generally use past occurrences for building the transitions and immediate outcomes of an internal model within the environment. The environment in reinforcement learning is described using mathematical frameworks called Markov Decision Processes (MDPs) (Littman, 2015). An MDP, mostly used in model-based reinforcement learning, comprises determinate environment states, possible actions in each state, a real-valued reward function, and a transition model. On the other hand, Model-free approaches utilize past occurrences to learn directly from the state/action values or policies for achieving optimal behavior but without estimating the model. The commonly used model-free RL methods include policy optimization and Q-learning. The policy optimization methods involve learning straight from the mapped state to action policy function but without value function. While, Q-learning consists of updating values of action in states; also known as Q-value. The updating value is fundamental for the Q-learning algorithm. Other algorithms used in the model-free approach are Deep Q-Networks, Deep Deterministic Policy Gradient, and many more (Ding et al., 2020).

Owing to its flexible component the reinforcement learning has been employed in numerous disciplines including control theory, game theory, information theory, simulation-based optimization, swarm intelligence, multiagent systems, statistics, and genetic algorithms among others. It is extensively used in building gameplay, robotics, autonomous vehicles, and many more; mostly where large volume simulated data is the input data. Reinforcement learning is used in the development of AlphaGo Zero. It was also used in the development of other games including ATARI games and Backgammon (Elfwing et al., 2018). Besides, the dialog agents (text, speech), and text summarization engines, designing optimal treatment policies in healthcare, and online stock trading are some other areas where reinforcement learning is employed as they can improve with time and from user interactions.

37.3.1.4 *Semisupervised learning*

The supervised and unsupervised learning are founded on the requirement the data must follow a predefined rule like labeled or with outcomes variables for supervised and unlabeled or without outcomes variables for unsupervised. But real-world data is of varied nature. Like in an image archive some images might be labeled or others might be unlabeled. Furthermore, hand-labeling of data by an ML engineer or a data scientist is expensive and time-consuming, particularly the big-data. These kinds of input data cannot be modeled using either supervised or unsupervised algorithms. The problems that sit in between supervised and unsupervised learning, that is, when big input data (X) and few outcome variables (Y) are available, the semisupervised learning algorithms are employed (Tu, 2019). Numerous ML challenges fall under this. The semisupervised ML paradigm comprises features of both supervised learnings (labeled training data) as well as unsupervised learning (absence of labeled training data). The amalgamation of unlabeled data with labeled data has demonstrated augmentation in learning accurateness (Goldberg, 2009).

The most common approaches of semisupervised learning are label propagation algorithms and semisupervised generative adversarial networks (GAN). There are three assumptions of the Semisupervised algorithm about the data, that is, continuity, cluster, and manifold assumption (Dligach et al., 2015). Continuity assumption reflects the assumption of the algorithm that the closer data points will probably have the same output variable. Cluster assumption reflects the assumption of the algorithm that if data points on clustering belong to the same clusters then they will probably share

an output variable. Manifold assumption reflects that the data lie in a much lower dimension than the input space and the algorithm is working on approximation (Cholaquidis et al., 2020). Speech analysis to label audio files, internet content classification (ranking relevance by search engine against a user query), and protein sequence classification among others are some of the areas where semisupervised learning has been successfully applied.

37.3.2 Artificial neural network

ANN or neural network is a subset of ML; they are trainable algorithm that uses a network like a topology. These networks also “learn” to execute tasks by studying examples, and not by explicitly. These are also known as connectionist systems as they vaguely mimic neural networks of brains (Okwu & Tartibu, 2021). A set of connected nodes are called “artificial neurons” that constitute an ANN model, which is similar to the neurons of biological brains. Akin to the synapses in the brain, each association of artificial neurons can pass on information called a “signal” from an individual neuron to another (Buscema et al., 2018). Artificial neuron after obtaining the signal, process and then traverses the signal to the next neuron. ANN was developed in the early 1940s by Warren McCulloch and Walter Pitts based on threshold logic; whose further development led to the launch of Perceptron in 1958 by Franck Rosenblatt. However, they were not widely used until backpropagation was created in 1969. The traditional ANN method bore with problems like diminishing gradients, and overfitting (Basheer & Hajmeer, 2000).

Typically, ANN implementations have numbers/value as a signal at the edges, that is, a connection between neurons but the output of each neuron is computed from the summation of its inputs by a nonlinear function. Both artificial neurons and edges carry an adjustable weight that increases or decreases during the learning process according to the signal/value strength at the edges. ANN sometimes also has a threshold against so that only those higher than the threshold are received as inputs. ANN is a layered network where every layer carries out different kinds of signal/value transformations. The three layers are Input, Hidden, and Output layer. The input layer comprises input nodes having the raw data. The number of input nodes indirectly linked with the explanatory variables count. The input layer performs data duplication which is then fed into the hidden layers of the network. The hidden layer contains hidden nodes that receive data from input nodes and perform actual processing using weighted connections or threshold logic. The number of hidden layers might be more than one. The output nodes in the output layer receive information either from the hidden or input layer following which predicted outcome will be returned as output. The activity of the output node is closely linked with the hidden nodes and the weights of the hidden/output nodes. Signals/values travel from the first layer (the input layer) to the last layer (the output layer), possibly subsequently travel through the numerous hidden layers (Basheer & Hajmeer, 2000; Buscema et al., 2018). The central aim of the ANN is to decipher solutions to the problems like a human brain. The two ANN topologies are FeedForward and Feedback. The flow of information is unidirectional in FeedForward. Every node/layer sends information to other nodes/layers but won't receive from them; due to the absence of feedback loops. They are used in recognition, pattern generation, and classification. Whilst feedback loops are permitted in Feedback topology which is mainly utilized for content-addressable memories (Yao, 1999). ANNs are capable of following several learning strategies like supervised, unsupervised, and reinforcement learning. ANNs have been employed in a variety of areas, including machine translation, computer vision, social network filtering, speech recognition, bioinformatics analysis, and medical diagnosis among others.

37.3.3 Deep learning

DL is the advanced version of ML, wherein the machines can learn from experience and gain skillfulness without human intervention. DL, also called hierarchical learning or deep structured learning, is inspired by ANN and also uses a layered architecture for data analysis. DL can solve more complex problems with a large number of features as massive parallelization is performed in it. The “deep” in “deep learning” refers to the possession of numerous (deep) layers in the neural networks that facilitate learning. The DL algorithm carries out a task repetitively and alternating it every time for the improved outcome, resembling the human learning process. Any problem that necessitates “thought” to solve can be figured out by DL; as it aids machines for solving complex using diverse, unstructured/unstructured, and inter-connected data (Lecun et al., 2015). Le Cun, Geoffrey Hinton, or Joshua Bengio are considered the fathers of DL; also received the prestigious Turing Award in 2018. The conceptual foundations and engineering advances laid by LeCun, Bengio, and Hinton aided by graphics processing unit (GPU) computers and massive data are key behind the DL and AI success. Backpropagation and Boltzmann Machines in 1983 by Hinton; Convolutional neural networks in the 1980s by LeCun; hidden Markov models in 1990s, GAN in 2010 by Bengio; and Improved backpropagation

algorithms by LeCun along with Improved convolutional neural networks in 2012 by Hinton laid the foundation of DL (Muthukrishnan et al., 2020).

Compared to ML, where an inaccurate prediction by algorithms is checked and accordingly adjusted for achieving correct prediction, the DL models can establish accuracy of prediction by themselves. Besides, the learning in DL can be supervised, or unsupervised, or semisupervised. The DL model, similar to the neural network, contains Input, Hidden, and Output layers. While ANN permits two hidden layers; DL allows around 150 hidden layers. Owing to the enormous number of layers, DL can perform more complex operations. The most popular types of DL algorithms are Convolution Neural Networks (CNN), Long Short-Term Memory Networks (LSTMs), and Recurrent Neural Networks (RNN) among others (Ahmad et al., 2019). Although none of them is perfect for every problem, some outdo others in specific tasks. CNN, also known as ConvNets, includes several layers and is largely used in object detection and image processing. Yann LeCun designed the first CNN in 1988, and named it LeNet, for distinguishing typescripts, that is, digits and Pin codes. The layers of CNN are Convolution Layer, Rectified Linear Unit (ReLU), Pooling Layer, and Fully Connected Layer. The convolution layer contains multiple filters for feature extraction and data convolution. ReLU layer operates on elements and generates rectified feature map as an output which is then fed into a pooling layer. The pooling layer executes dimensionality reduction of the map and transforms it into a single, continuous linear vector. The fully connected layer classifies and identifies the linear vector image (Tran et al., 2015; Yang et al., 2015). Identification and processing of satellite or medical images, forecasting time series, and detecting anomalies among others are some of the widespread applications of CNN. Other DL algorithms such as LSTMs and RNNs learn, memorize, and recalls long-term dependencies and past information. RNNs form a directed cycle, while LSTMs form a chain-like structure. LSTMs retain information of past inputs and update values based on relevance; therefore they are used for time-series prediction, music composition, speech recognition, and pharmaceutical development. RNNs allow the outputs at time $t-1$ or t to be fed as inputs at time t or $t+1$ based on directed cycles architecture due to its internal memory. RNNs are commonly employed in time-series analysis, handwriting recognition, image captioning, NLP, and machine translation (Aggarwal & Murty, 2021; Malhotra et al., 2015).

Besides CNN, LSTN, and RNN, several other DL algorithms that can be used in developing DL models are Radial Basis Function Networks, GAN, Restricted Boltzmann Machines, Multilayer Perceptrons, and Deep Belief Networks (DBNs) (Schmidhuber, 2015). Although the majority of DL models' attributes feature mining to its layered architecture, they are also employed for propositional formulas or layer-wise organized variables in DBN, deep generative models, and deep Boltzmann machines. Advances in DL turned it a popular choice for AI application as in computer and machine vision, speech and audio recognition, NLP, social network filtering, machine translation, bioinformatics, drug design, and board game programs, among others have yielded comparable results, even surpassing human expert performance in some cases (Ahmad et al., 2019).

37.4 Technological advancements in artificial intelligence

AI is not a miracle; in many cases, its functions depend on a physical device that incorporates various algorithms of AI. Like a robot without its body is of no use, AI software also often needs a shell, that is, hardware to be productive. In the past, there was a heated argument about the higher significance of hardware or software for computer advancements. Eventually, as computer hardware standardized, the intervening decades beginning to focus on software development; concluding that both hardware and software work hand in hand. Similarly, the advancements in AI are closely linked with computer algorithms, hardware, and software. The current AI advancements are similar to computer technology 1970s and 80s. Therefore the argument about the higher significance of hardware or software for AI advancement is at the peak and the chip research is again going through the same transition.

To understand the link between hardware and software, one has to grasp technological layers or stack in the computer application architecture. Technological stack, a list of all the technology services used to build and run one single application, of AI can be divided into five layers, that is, Hardware, Interface, Platform, Training, and Services (Fig. 37.2). Hardware is at the bottom with no direct user interaction and services at the top with full functioning, ready for the users. Hardware is comprised of the accelerator and the head node that deals with the highly parallel operation and computation among accelerators as required by AI, respectively. An accelerator is a silicon-based chip that is used for computer memory, storage, logic processing, and networking. The next layer in the technological layer is Interface, which facilitates the communication between software and underlying hardware. The platform layer, on the top of the Interface, deals with the software packages, rules, approaches to be applied to the data for the analysis. It comprises four sublayers, that is, Framework, Algorithm, Architecture, and Methods. The Framework is where the software is used to define and invoke algorithms on hardware through the Interface layer. The algorithmic rules with weights to utilized in the model training

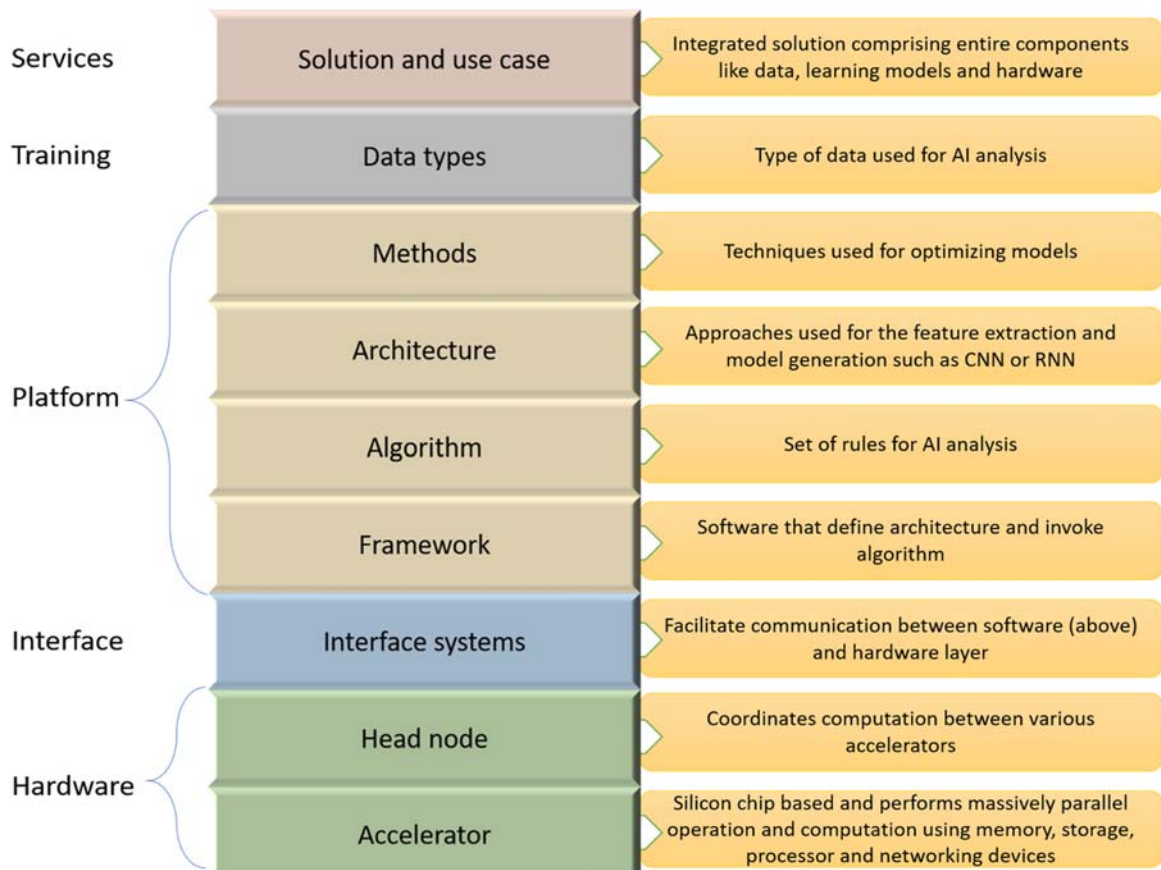


FIGURE 37.2 Artificial intelligence (AI) technological stack layers. The technological stack provides a list of all the technology services used to build and run one single application. AI technological stack is divided into five layers, that is, hardware, interface, platform, training, and services.

and optimization and approach that to be used on the data for feature extraction are mainly performed by Algorithm and Architecture sublayers; whereas Methods sublayer primarily dealt with the techniques for optimizing weights given to models respectively. Training and Services, of the technology stack, respectively deals with the data types of data for analysis and final integrated solution that includes training data, model, hardware, and every other component (for example voice recognition system, handwriting recognition system) (Quigley et al., 2007).

This section will focus on the technological advancements in both hardware and software that have helped AI to achieve several of its accomplishments.

37.4.1 Hardware

In recent years, AI has witnessed immense progress with the advent of deep neural networks and surpassed humans in the number of cognitive tasks. With time AI is becoming more sophisticated and demanding more computation power for achieving its full potential. Newer hardware, designed for AI, is meant to accelerate the training and performance of AI models with reduced power consumption. Essentially, the AI hardware consists of computer memory, storage, logic, and networking. Computer memory primarily helps in the temporary storage of data and instructions during processing, for example, dynamic RAM or DRAM, while computer storage helps in the long-term storage of large datasets; typically using Not AND or NAND type of devices. The computer logic involves its logical function, that is, optimization and calculation of neural network operations; mainly performed by processors and typical logic devices are central processing unit (CPU), GPU, application-specific integrated circuit (ASIC), and field-programmable gate array (FPGA). The computer networking utilized switches, routers, or other networking components (Does, 2018).

The primary AI hardware is called accelerators, that is, silicon chip-based microprocessors or microchips that facilitates faster AI processing due to the presence of multicores, novel dataflow architectures low-precision arithmetic, or in-memory computing (IMC) capabilities. and parallel task completion (Welser et al., 2019). Additionally, the

algorithmic complexity of DL poses high computation and memory demands which are challenging to the hardware platforms. This section will discuss AI hardware.

37.4.1.1 Processor

Processors, invented in 1937 by Marjian Hoff, are the logical circuits that handle the computer instructions within seconds and their speed is measured in terms of megahertz. The four main primary tasks of the processor are fetching, decoding, executing, and write back the instructions. They are called the brain of the system including computers, smartphones, robots, embedded systems, etc. The processor receives input devices instruction after processing sends the output to the output devices. The Arithmetic Logic Unit (ALU) and Control Unit (CU) are the two components of the processors. The ALU executes all mathematical operations and the CU operates like traffic police by managing the instructions command. Typically, the CPU is the main processor in almost every electronics device surrounded by the microcontroller. The operations of a processor are inherently constrained due to the sequential execution of instructions. To address this, one part of hardware designing is also focused on multicores. Currently, every CPU includes multiple processing cores to accomplish the execution of multiple tasks simultaneously. Even though “cores” are confined in one physical unit, they are independent processors. For instance, processors with two cores are dual-core, with four cores are quad-core, and eight or ten are known as octa-core or deca-core (Chen, 2016b). The processor for AI includes CPU, GPU, FPGA, an ASIC (Mittal, 2020; Momose et al., 2020).

Although the CPU is the main processor, more often not it is not an AI accelerator. In the beginning, CPU was mostly employed for AI-related studies but as computational demands of AI, increased CPU efficiency strained owing to its physical limits in clock speed and heat dissipation, and development of newer hardware becomes a necessity. The GPU, FPGA, and ASICs are known AI accelerators. The AI computations need only linear algebra for parallel processing and GPUs are specifically designed for massively parallel operations required for graphics rendering to achieve smooth video display. The performance of the CPU is enhanced with the aid of GPU; as it can take the computationally-intensive load of the CPU. GPUs are first built by NVIDIA and are an assemblage of hundreds to thousands of cores in parallel. Traditional CPUs generally took weeks for prediction whilst GPUs may only take days or hours for the same task (Lee et al., 2010). In 2009, Andrew Ng’s group illustrated GPU’s capability for large-scale DL. The group reproduced the 11 billion network connections of the Google X project with 16 computers powered by 64 GPUs whereas originally 1000 computers with 16,000 CPUs are used. The replicated project did not run significantly faster or performed better, but it demonstrated that 64 GPUs is equivalent to 16,000 CPUs. Ever since GPUs are a popular choice for AI-related training and inference. Newly developed GPUs such as Nvidia NVLink and Tensor cores have enhanced connective capability for AI dataflow and are neural network-specific hardware (Böhm et al., 2009; Coates et al., 2009). FPGA is a type of programmable logic device that can be easily be configured and optimized for the latest AI algorithms. As DL frameworks are still developing, it is difficult to design customized hardware for it, but easily reconfigurable devices, FPGA, are simpler to customize and can be evolved besides frameworks and software. FPGA chips are used to accelerate real-time AI inference under project brainwave launched in 2017 by Microsoft. Intel acquired Altera in 2015 to integrate FPGAs in the CPUs server for accelerating AI as well as performing general-purpose tasks (Mittal, 2020).

Although GPUs and FPGAs execution for AI-related tasks are far better than CPUs, it is believed that with a more specific design that efficiency can be improved ten times. ASICs are integrated chips that are tailor-made for a specific purpose or application. These accelerators through optimizing memory usage and lower precision arithmetic accelerate the calculation as well as increase the throughput of computation. ASICs are often utilized in big data products such as consumer or business products or cell phones or other similar applications (Mittal, 2020; Momose et al., 2020). Currently, GPUs are the ultramodern hardware in the machine and DL. Owing to their good performance in parallel computing to the thousands of cores. Nevertheless, the ever-growing deep neural network complexity has directed the search for advances in processing efficiency. The AI hardware researchers are investigating novel algorithms, architectures, devices, and approaches like quantum computing, IMC, and approximate computing for enabling AI workloads and the shift from Narrow AI (ANI) to Broad AI (AGI or ASI).

Recently developed new generation of hardware architectures are optimized for AI workloads; for instance, Qualcomm’s Snapdragon 888 processor is a computer vision and AI driving force. Currently, popular AI-specific hardware includes:

1. Tensor Processing Unit (TPU) is developed by Google. It is an ASIC type of AI accelerator specifically designed for neural networks and ML solutions. TPU also empowers the Google products like Assistant, Search, Translate, Photos, and Gmail. It also offers cloud TPU for executing other ML teams’ solutions.

2. EyeQ is a system-on-chip (SoC) device designed by Mobileye consisting of 32-bit ASICs microprocessor, memory blocks, network circuits, etc. EyeQ is optimized for complex and computationally heavy signal-processing, computer-vision, and ML/DL tasks while using low power. Currently, it has been incorporated into assisted-driving technologies, even in fully-autonomous (Level 5) vehicles, by more than twenty-seven car manufacturers (Lloyd, 2015).
3. Nervana Neural Network Processor-T 1000, developed by Intel is a discrete accelerator designed specifically for the ever-increasing AI complexity and scale of inference (Hickmann et al., 2020).
4. The Intel Movidius Myriad is a vision processor unit SoC type of device containing programmable processors, dedicated and configurable image and vision accelerators. It is specifically designed for on-device computer vision and neural network applications. The device presents top-tier performance per watt for demanding workloads in AI inference (Bakshi & Johnsson, 2020).
5. Epiphany V is developed by Adapteva with a 1,024-core processor chip aimed for real-time ML in the form of image processing and autonomous driving (Varghese et al., 2017).
6. Akida neural processor is developed by BrainChip using the latest neuromorphic computing. This DL accelerator is low power, inexpensive, and follows incremental learning. Earlier the distribution was with license; but now it is freely available on their site, Github, and Python. Currently, BrainChip is developing a software toolkit, the Akida Development Environment (Lorenc et al., 1973).

Besides, the interest and investment in FGPA for AI training and inference is also increasing. NVIDIA's latest Tesla V100 and NASDAQ: NVDA GPUs are DL-focused and are better suited for AI applications. The start-ups like Mythic, Wave, and Graphcore Computing are also working toward faster and cheaper AI training and inferencing. One aspect involved in hardware advancement is closely linked to DL growth both in the framework as well as market; therefore everyone ensuring that their chipsets are an advantage, not a pitfall.

37.4.1.2 Memory device

Memory devices are required to store information for immediate use. Three major processes involve memory: encoding, storage, and retrieval. The computing layers of DL or ANN models require a large amount of data to pass rapidly into thousands of cores for processing; that's why AI applications have high memory-bandwidth requirements, that is, the rate of reading or storing data into memory by a processor. For example, a model trained for identifying the cat image requires colors, contours, textures among others to reside on memory during the recognition process. The dynamic random-access memory (DRAM) is most widely used in AI for storing input data, weighing model parameters, and performing other functions during both inference and training. DRAM was invented by Robert Dennard while working in the Research Division of IBM; the "dynamic" refers to the constant refreshing of the charge on each capacitor in DRAM. The invention ultimately led to the formation of a single chip that can hold billions of RAM cells in modern computers. The first commercial DRAM, that is, Intel 1103 was launched by Intel in the 1970s. Since then, numerous versions of DRAM have been developed that are keeping up with the latest accelerators and algorithms (Upadhyay et al., 2019). Still, currently, the development in-memory infrastructure-related field is the least among the AI hardware and technologies. The major contributing factor behind its slower development might be the rapid advent in the micro-processor chips/accelerator and improved efficiencies in algorithm design such as reduced bit precision and their ability to be par at recent advancements. Existing memory is augmented for CPUs, but new architectures are being explored. Since the launch of DRAM, various advancements like Double Data Rate (DDR), Graphics Double Data Rate (GDDR), and Low power DDR (LPDDR) among others have been achieved. Nowadays, GDDR is one of the popular choices due to its close integration with the processor for applications with a high bandwidth demand (Kim et al., 2016). Nonetheless, a large GPU can only be surrounded by twelve GDDR chips; therefore there is still a limit in the bandwidth it can offer. Other memory solutions include High-bandwidth memory (HBM) and On-chip memory that is also closely integrated with the GPU (Jun et al., 2017).

The newer designs and architectures for the memory mainly involve through-silicon-vias (TSVs), that is, vertical fast interconnections on-chip that enables 3D memory stack accessibility to the processor and resulting in higher bandwidth memory. HBM a 3D equivalent of GDDR, Hybrid memory cube (HMC), developed by Micron was a proposed 3D equivalent of DDR, and Wide I/O is a 3D equivalent to LPDDR memories in SoCs devices for improving AI processing. Owing to the higher number of parallel interconnects, the power consumption per bit in these newer 3D architectures is three times lower. With HBM technology large datasets can be processed rapidly while reduced power requirements. HBM is currently the memory solution of choice for Google and Nvidia, even with its three times expenses against traditional DRAM per gigabyte (Tran et al., 2015; Yang et al., 2015). HMC primarily focused on the

high total system memory capacity and easy plugging into a server as memory stacks, similar to DDR memories (RAM). It provided a loose integration and called as far-memory, but HMC was canceled. The Wide I/O is an example of extreme integration for achieving the lowest possible power consumption. The memories are integrated directly on top of SoCs and are connected directly to the CPU using TSVs. However, this integration requires TSVs in the SoC, which consumes a lot of precious logic area, and thus is pricey (Hansson et al., 2014; Kim et al., 2016). This is perhaps the biggest reason behind the absence of its implementation in any commercial products yet.

The AI-related data computation, storage, and access in DRAM or other outside memory sources can take 100 times more time than if the memory is located on the same chip. These can be obtained through Non-von Neumann computing architectures like IMC/ Processing-in-memory (Bavikadi et al., 2020; Gauchi et al., 2019). Unlike von Neumann architecture where memory and processor are separate units (e.g., personal computer), Non-von Neumann computing architectures contain both on the same units; this, in turn, targets the conceptual constraint of traditional computing systems. On-chip memory is developed using IMC architectures so that data does not need to be constantly exchanged between RAM and the processor (Gauchi et al., 2019). Through increasing on-chip memory the AI speed can be enhanced as the time taken for data exchanged will be minimized. For instance, the ASIC type of Google's TPU processor comprises sufficient memory to store an entire model on the chip. The Graphcore are working in the direction and further taking it to a level about 1000 times than found on a typical GPU, through a novel architecture. The expense of on-chip memory is still excessive for most applications (Jia et al., 2019; Kacher et al., 2020). In 2018, IBM announced an in-memory-based architecture for processing and modeled according to the synaptic network of the brain to accelerate DL. Every solution has a different target area for achieving better performance like DDR and HMC focus is capacity and flexible integration, LPDDR and Wide I/O focus on the lowest possible power consumption, and GDDR and HBM focus on the highest bandwidth.

37.4.1.3 Storage device

AI techniques, both ML or DL, require data and storage architecture differently from traditional workloads; as they produce large volumes of data (~ 80 exabytes annually and projected to be 845 exabytes by 2025). Also, the amount of data used in AI training is growing, hence a further surge in storage demands. From 2017 to 2025, the estimated annual growth of storage is 25%–30%. Unlike conventional storage solutions that follow a one-size-fits-all approach, AI solutions must adjust according to changing needs. Both AI training and inference models have high storage demands and store massive volumes of data during algorithm refinement, but overall, it is higher for AI training compared to inference. Traditionally, the data storage is performed by Nonvolatile memory (NVM) using semiconductor memory chips where stored data is retained even after power disconnection. For instance, read-only memory, magnetic storage devices (like magnetic tape and hard disk drives), optical disks, and flash memory (solid-state drives and NAND flash) among others. Ideally, AI needs a storage device that is fast like static random-access memory, has storage capacity similar to DRAM or Flash (Chen, 2016a; Xue et al., 2011), and has low power dissipation. None of the current technology fulfills these demands, resulting in a “memory bottleneck” reflecting severe limitations in the performance of AI applications.

The recently developed NVMs fall between traditional memory (DRAM) and traditional storage (NAND flash) but have a higher density than traditional memory, better performance than traditional storage, and lower power usage than both. The multiple NVM technologies such as magnetoresistive random-access memory (MRAM), Resistive random-access memory (ReRAM), and Phase-change memory (PCM), differing in terms of memory access time and cost are in various development stages. MRAM has data retention for more than five-year, high endurance, and the lowest latency for reading and writes function. However, due to its limited scaling capacity, it is an expensive substitute for caches and not a long-term data-retention solution (Tsymbal et al., 2019). Whereas, ReRAM has potential in vertical scaling and advantageous pricing, but it has slower latency and reduced endurance (Mittal, 2018). PCM fits in between the two, but endurance and error rate needs to be addressed for its widespread adoption (Ambrogio et al., 2020). The development of antiferromagnetic (AFM) materials could potentially aid in data-centric computing that requires ever-increasing power, storage, and speed. The AFM-based device is the smallest, operates with a record-low electrical current for data writing, and has electrons in antiparallel alignment that behave like tiny magnets due to spin. In AFM materials, the constant electric current is not needed and data cannot be erased by external magnetic fields due to dense packing which will not interact with magnetic fields. Besides, AFM-based devices are also very secure and easy to scale down to small dimensions (Ha et al., 2004). The FlashBlade is developed and optimized by Pure Storage for the fulfilling storage needs of the entire ML workflow. It has a datacenter infrastructure, high-performance file and object storage, scalable elastic storage, and massive throughput (~ 75 GB/s from a cluster of devices) for any access pattern, sequential or random (Stalzer, 2012). Similar to GPUs, it is the first in the storage industry that is immensely parallel in architecture.

37.4.2 Software

The tremendous progress of AI in recent years with the aid of DL and ML has made AI systems more sophisticated and demanding computation power from hardware. New hardware, as well as software, are being specifically devised for accelerating AI training and performance. The software component of AI can be divided into AI platform and software solution, where AI platform is defined as hardware architecture or software framework (including application frameworks) that allows execution of software or software solution over it.

37.4.2.1 Artificial intelligence platform

AI platforms facilitate software execution and application. It stimulates the cognitive function of minds to perform problem-solving, learning, reasoning, social intelligence as well as general intelligence. AI platforms can be divided into weak AI/ narrow AI (used for a specific task) or strong AI (or ASI) (find solutions for unfamiliar tasks). Recently popular cloud infrastructure offers scalability as well as resource access for the implementation of complex AI and ML solutions. It is vital to regulate platform as a service (PaaS) and software as a service (SaaS) while launching AI solutions for best performance (Beimborn et al., 2011; Tsai et al., 2014). Microsoft Azure, Google AI Platform, TensorFlow, Amazon AI Services, Rainbird, Wipro HOLMES, Infosys Nia, are some of the top AI Platforms.

37.4.2.1.1 Google artificial intelligence platform

It is a simple, quick, and cost-effective platform available for building customized ML projects and applications from conceptualization to production to deployment. It also supports an open-source platform, Kubeflow for the construction of portable ML pipelines to run on-premises or on Google Cloud. AI Technologies such as TensorFlow, TPUs, and TFX tools are accessible through the Google AI platform. Additionally, libraries of Prediction and RESTful API for successful integration of search engines are available in popular languages, such as Python, JavaScript, and DotNET. It is primarily used for Cloud-based ML, Spam Detection, Customer Sentiment Analysis, Recommendation Systems, and Purchase Prediction among others (Hlavac et al., 2004).

37.4.2.1.2 TensorFlow

It is a simple, visual, and open-source software platform created by Google Brains' team to implement ML and DL Neural Networks for numerical computation; it provides ML capabilities for different programming environments and visual interface that relies mostly on graphs and data visualizations. It was the first highly accessible ML platform and resulting in the widespread implementation of ML in training models, JavaScript, and bringing ML to mobile, and IoT devices. Besides, the Keras library is available for Neural Networks programming (Abadi et al., 2016; Nelli & Nelli, 2018). Currently, it works with both GPU and TPU; it is flexible enough to permit users to use one or more GPUs or CPUs in a mobile, or desktop, or server with a single API (Abadi, 2016; Ketkar & Ketkar, 2017).

37.4.2.1.3 Amazon artificial intelligence services

Amazon Web Services (AWS) is one of the most widely adopted cloud platforms. It offers a variety of services, features, infrastructure, and emerging technologies for computing, storage, ML and AI-based data analytics, and the Internet of Things. AWS provides the widest variety of specifically built databases for different types of applications and offers them for best cost and performance. Its flexible and secure cloud computing environment is designed for fulfilling security requirements for global banks, military, and other high-sensitivity organizations. In 2014, AWS Lambda, a serverless computing space, was launched to allow developers to execute their code without provisioning or managing servers. Amazon SageMaker, another AWS service, is a ML service that lets everyone employ ML without prior understanding (Muni & Hansen, 2005; Varia & Mathew, 2014).

37.4.2.1.4 Microsoft Azure

It is mobile-enabled, cloud-based advanced analytics that supports every operating system, language, tool, and framework developed to simplify ML for businesses. Business users can model through algorithms from Bing, Xbox, R, or Python (both customized and noncustomized R or Python code) and can be uploaded as a web service or in the product Gallery or into the ML Marketplace (Schwichtenberg & Schwichtenberg, 2020). Some of the areas where it has been successfully employed are Monitoring, Digital Marketing, Business Intelligence, E-Commerce, Big Data and Analytics, Digital Media, Internet of Things, High-Performance Computing, Gaming, and Blockchain, among others.

37.4.2.1.5 Rainbird

It is an award-winning AI platform that enables the building of a decision-making system that increases efficiency and quality of customer interaction by combining current human-business knowledge. It automates work and provides consultative systems for enterprises. It has been utilized for producing Analytics & Insights, RBLang, Visual User Interface, Smart Data Import, Controlled Learning Algorithms, and NLP among others (Hlavac et al., 2004).

37.4.2.1.6 Infosys Nia

It is a knowledge-based AI platform that combines ML with deep insights to create automation and innovation by simplifying the continuous renovation of core processes within an organization. Nia also enables businesses to bring new, delightful user experiences leveraging state-of-the-art technology. The Infosys Nia platform belongs to the Infosys Aikido framework which primarily deals with the order-to-activation process transformation asset efficiency and incident automation. Jointly with AiKiDo service, it offers reduced maintenance costs for both physical and digital assets. Major platforms of Infosys are information platform, automation platform, knowledge platform, and aikido framework (Hlavac et al., 2004; Wilamowski & Irwin, 2016).

37.4.2.1.7 Wipro HOLMES

It is developed using ML, DL, NLP, genetic algorithms, pattern recognition semantic ontologies, and knowledge modeling technologies to deliver improved cognitive experience and productivity along with accelerating the process through automation and autonomous abilities. This AI platform provides a wide variety of cognitive computing services such as predictive systems, knowledge virtualization, virtual agents, visual computing applications, cognitive process automation, robotics, and drones (Tarafdar & Beath, 2018).

Other AI platforms include H2O, Petuum, Polyaxon, DataRobot, NeuralDesigner, PredictionIO, Dialogflow, MindMeld, Premonition, Ayasdi, Meya, KAI, Receptiviti, Watson Studio, Vital A.I, Wit, Lumiata, and Infrd among others. Additionally, various APIs are available for the development of software targeting particularly for Robotic Automation (UIPath, BluePrism, and Pega Platform), Chatbot Software (Engati, Chatbot, ManyChat, and FreshChat), Facial Recognition (Deep Vision AI, FaceFirst, Trueface, and Amazon Rekognition) among others.

37.4.2.2 Artificial intelligence solution

AI Software Solutions can provide a sustainable and cohesive AI-driven ecosystem by conceptualizes and implementing data-driven decision making that can assist in informed decision making, by identifying growth hacking opportunities, trends, and anomalies in operational processes, including risk analytics, predictive maintenance, operational forecasting, and demand prediction. Marketing & Sales, Customer Service, Mobile Application Solutions, IoT Solutions, customer service, and Predictive Analytics Solutions are some of the areas where AI software is available in abundance. Table 37.2 listed a few of the AI software.

37.4.2.3 Big data

Although it is not technical, it also played a role in the implementation of AI-related applications; as the data can be without information, but information cannot be without data. The concept of big data, even without the term “big data,” has been around since the 1990s and by the time term was coined in 2005 by Roger Mougaldas from O’Reilly Media, the massive amount of data has been already accumulated. The rise of big data has enabled the emergence of AI cloud and on-edge devices which is altering the computing, networking, and data storage industries fundamentally. Big data is the power behind AI; as diversity in big data is making ML and DL applications do what they were designed to do, that is, develop and improve a skill. The amount of data available for the AI is unequivocally linked to its learning and improvement of pattern recognition capabilities (Haenlein & Kaplan, 2019; O’Leary, 2013).

37.5 Application of artificial intelligence

While the promise of AI is not yet fully realized, and according to AI scientists and experts, it’s still in its infancy. Nevertheless, its endless potential and human-like capabilities look very promising. Currently, AI applications have already been popular in various fields like customer interactions management, healthcare, chatbots, computer vision among others. AI-driven technologies are proving to be beneficial in many fields including agriculture through crop and soil monitoring, weather forecasting, predictive agricultural analytics, and markets. The cloud computing

TABLE 37.2 List of artificial intelligence software.

AI Tools	Functionality	Supported OS/ Languages/ Platform	Best Feature	Price
Content DNA Platform	Machine Learning/ Computer Vision.	Suits both Cloud and On-premises deployment models.	Unsupervised Machine learning. Training on your data.	One-time fee.
Google Cloud Machine Learning Engine	Machine Learning	GCP Console	Trains model on your data. Deploy it. You can manage it.	Per hour per training unit costs: US: \$0.49; Europe: \$0.54; Asia Pacific: \$0.54
Azure Machine Learning Studio	Machine Learning	Browser based	Model will get deployed as a web service.	Free
TensorFlow	Machine Learning	Desktops, Clusters, Mobile, Edge devices, CPUs, GPUs, & TPUs.	It is for everyone from beginners to experts.	Free
H ₂ O AI	Machine Learning	Distributed in-memory. Programming. Languages: R & Python.	AutoML functionality included.	Free
Cortana	Virtual Assistant	Windows, iOS, Android, and Xbox OS. Supported Languages: English, Portuguese, French, German, Italian, Spanish, Chinese, and Japanese.	It can perform so many tasks from setting reminders to switching on the lights.	Free
IBM Watson	Question-answering system.	SUSE Linux Enterprise Server 11 OS Apache Hadoop framework.	It learns lot from small data.	Free
Salesforce Einstein	CRM system	Cloud based.	No need for managing models and data preparation.	Contact them for pricing details
Infosys Nia	Machine Learning Chatbot.	Supported devices: Windows, Mac, & Web based.	It provides three components, i.e., Data platform, Knowledge platform, and automation platform.	Contact them for pricing details.
Amazon Alexa	Virtual Assistant	OS: Fire OS, iOS, & Android. Supported Languages: English, French, German, Japanese, Italian, and Spanish.	It can be connected to devices like Camera, lights, and entertainment systems.	Free with some amazon devices or services.
Google Assistant	Virtual Assistant	OS: Android, iOS, and KaiOS. Supported Languages: English, Hindi, Indonesian, French, German, Italian, Japanese, Korean, Portuguese, Spanish, Dutch, Russian, and Swedish.	Supports two-way conversation.	Free
PaleBlue	VR Simulations, creating Virtual Reality, Augmented Reality, and 3D simulators	iOS, Android, Windows, Mac, & Web based, Cloud, SaaS	PaleBlue is the leading provider of VR, AR, & 3D simulators for the real world. PaleBlue digital solutions help its clients to intensify training, streamline workflow, & improve safety worldwide!	Free Trial
BIRD Analytics	Machine Learning	iOS, Android, Windows, Mac, & Web based, Cloud, SaaS	Healthcare, Manufacturing, Financial Services, Insurance, Automotive, and Retail	\$8.00/month/user

(Continued)

TABLE 37.2 (Continued)

BAAR	Machine Learning	Installed – Windows, Web-Based, Cloud, SaaS	Automated Workflows, Computer Vision Capabilities, Reporting and Analytics, Low Code Platform, Robotic Process Automation, Industry-Specific Solutions	Not provided by vendor
G6GFINDR System	query-based system and Natural Language Processing (NLP)	Windows	Provide in depth search for info/ meta-data on the artificial intelligence and bioinformatics software fields. This online Artificial Intelligence system offers Chatbot, For Healthcare, Predictive Analytics, Process/ Workflow Automation, Virtual Personal Assistant (VPA) at one place.	Free Trial

infrastructures with the use of data ecosystems, IoT devices, and AI enables the development of digital agriculture that can in turn strengthen the farmers through smart farming, irrigation, fertilizer application, and harvesting among others (Murase, 2000; Zhao, 2020). In this section, the AI applications in agriculture, and their related area are listed (Fig. 37.3).

37.5.1 Agriculture/farming

As, agriculture is also a significant contributor to the country's economy and frequently affected by challenges such as ever-increasing population and food security, climate fluctuations, herbicide resistance, pollution, soil deforestation among others; therefore requires novel strategies for augmented crop yield. AI is gradually providing solutions to the several challenges faced in agricultural operations such as disease detection, crop phenotyping, yield monitoring, weather forecasting, irrigation management among others, and rising as a part of the industry's technological evolution. Broadly, the application of AI in agriculture can also be termed as precision agriculture or farming which is focused on soil, weather, and crop conditions (Khattab et al., 2019). With the aid of sensors and technological advances such as robots, satellites, GPS, and drones the valuable data on crop growth, soil characteristics, and weather conditions is obtained that can further detect hidden knowledge about agriculture production.

In this section, the implementation and applications of AI in providing next-generation agriculture tools including agriculture robots for crop health and soil monitoring along with predictive analytics is discussed.

37.5.1.1 Field mapping

It the key component of precision farming where inter and intra-field variability in the crop is observed measured and then used for the development of better farm management. It involves monitoring and evaluating the exact geometry of agricultural entities like fields or ponds with their precise perimeter and location, that is, local geological data. The aggregated geological data of agricultural areas are useful for monitoring, crowdsourcing, and farm management (Fritz et al., 2015). *MapIT*, a crowdsourcing tool developed for collecting geographic information of small objects and agricultural areas. The tool requires firstly the snapping picture of the target field followed by its outlining by the user which is then simplified by an inbuilt Douglas-Peucker algorithm. Jointly with the data from the built-in internal sensors, the distance of the object from the camera, and GPS location, the coordinates in the photo are projected to obtain a geological object of original geometry (Schmid et al., 2013).

37.5.1.2 Yield monitoring

Yield monitoring or mapping is a significant facet of precision agriculture that assists farmers in making educated decisions by providing ample information about their fields. Yield monitoring or mapping refers to the process of georeferenced data collection with the aid of farm equipment such as drones, tractors, or harvester along with the information



FIGURE 37.3 Application of artificial intelligence (AI). AI-driven technologies are proving to be beneficial in many fields including precision farming, field mapping, yield monitoring, irrigation management, agriculture robot and drones, crop scouting, disease detection and diagnosis, crop phenotyping, soil management, nutrient monitoring, smart greenhouse management, weather tracking and forecasting, system biology, advisory services.

including grain yield, moisture levels, and soil properties, among other during crop harvesting. Monitoring involves feeding of harvested grain into the elevator for sensing grain moisture followed by their transfer to holding tank for sensing grain yield and then displayed on the screen. The information obtained is then georeferenced to the field and the associated field data can further help farmers in assessing things such as when to sow, fertilize or harvest, the effects of weather, and much more (Magalhães & Cerri, 2007). Numerous methods, using a range of sensors and imaging techniques have been developed for data collection; whilst, computer vision and DL algorithms are employed for data processing (Khaki & Wang, 2019). Typically, at least five years of yield maps are essential to avoid reaching conclusions that are affected by the unpredictable factors of a particular year. Plantix, a DL application developed by Berlin-based agricultural start-up PEAT, detects potential nutrient deficiencies and defects in soil. Liao et. al. used spatiotemporal fusion of MODIS and Landsat-8 data to estimate yield phenology and biomass for soybean and corn (Liao et al., 2019).

37.5.1.3 Irrigation management

Irrigation management involves fulfilling the water requirement of crops through the management of time and water application without wasting any water, soil, plant nutrients, or energy. Time and again newer methods of irrigation have been introduced to reduced flood irrigation (Dolci, 2017). Currently, drip irrigation with embedded systems is the most prominently used in precision agriculture where water usage is reduced by exploiting parameters such as soil, pest, wind speed, solar radiation, humidity, plant density among others. Devices such as fertility meter and pH meter are installed in the field to evaluate soil fertility via evaluation of primary ingredients of soil such as potassium, phosphorus, and nitrogen. Besides, automated farm irrigators and microcontroller controls drip irrigation through irrigator pumps and wireless technology. Moreover, machine-to-machine technology (M2M) is being developed for communication and data sharing amongst each other via cloud or main network of agricultural field. An AI-based robot is developed for estimation of moisture and temperature with Arduino and raspberry pi3; and an AI-based strategy for estimation of drip tape irrigation based on ANN, least-square support vector machine, neurofuzzy c-Means clustering (NF-FCM), and neurofuzzy subclustering, are developed (Seyedzadeh et al., 2020; Shekhar et al., 2017). Similarly, an automated irrigation system wherein output of cameras and different sensors are used for detection of soil moisture, pressure, and temperature are shared over the network for better irrigation management.

37.5.1.4 Crop scouting

Crop scouting refers to the process of precise crop performance and pest pressure (infestations and disease) assessment for estimating economic risk as well as determining the effectiveness of potential intervention strategies. Usually, scouting is incorporated in Integrated pest management; but with the aid of AI, built-in sensors, specialized field instruments, and handheld computers with GPS help geotagging of crops and their issues (Geng & Dong, 2017; Kalischuk et al., 2019). Geotagging aids in their visualization on an aerial map that can further help farmers in making site-specific treatment decisions. DL CNN-based framework, that is, MAESTRO and automation spotted grasshopper and *Drosophila suzukii* using RGB images and has shown potential for UAV-based monitoring (Roosjen et al., 2020). The DL CNN and ML are also employed for differentiating pest-damaged and healthy wheat grains as well as for investigating the spatiotemporal spread of *Tuta absoluta*, a pest of tomato in South and Southeast Asia (Mkonyi et al., 2020).

37.5.1.5 Disease detection and diagnosis

Plant diseases adversely affect the yield and quality of the crop and their rate of progression depends on the existing crop condition and its susceptibility. The plant's diseases often display morphological changes such as colored spots and streaks in leaves, stems, and seeds whose timely detection might prevent economic losses faced by farmers (Cruz et al., 2019; Pathan et al., 2020). Several AI technologies including DL, CNN, K-Means clustering method, SVM among others have been employed for the detection of diseased and healthy crops with varying degrees of accurateness in models. The developed models or tools can be used for early, instant detection, classification, and diagnosis of plant diseases which could be further expanded to support an integrated plant disease identification system under real cultivation conditions.

37.5.1.6 Agriculture robot and drones

The development and programming of autonomous robots for the handling of significant agriculture tasks is one of the high valued applications of AI. Automation and Robots are increasing precision as well as managing the farms by carrying out many operations such as irrigation, weeding, harvesting, and safeguarding the farms among others (Yahya, 2018). See and spray, a robot developed by Blue River Technology, employs computer vision to examine and accurately spray herbicide only on weeds of cotton plants; effectively reduces the chemical usage as well as targets potentially resistant weeds. On the other hand, the Harvest CROO Robotics robot aids strawberry farmers in their picking and packing of their crops. Similarly, drone technology developed by SkySquirrel Technologies Inc. also utilizes computer vision for examining crop health and has been successfully applied in reporting the health of the vineyard, particularly the condition of grapevine leaves.

37.5.1.7 Crop phenotyping

Phenotyping refers to all the features of an organism which include size, shape, color, biochemical properties, and behavioral properties among others that are the outcome of the interaction of genotype (total genetic inheritance) with the environment. Current phenotyping technologies have applied imaging techniques in combination with computer vision for plant phenotyping that can be applied in agriculture and crop science (Dolci, 2017).

37.5.1.8 Soil management

Soil monitoring involves measuring soil temperature, water potential, oxygen levels, NPK, and volumetric water content using sensors and IoT devices to maximize yield, reduce disease and optimize resources. The IoT application in agriculture is known as Smart Agriculture (or Smart Farming), and IoT is the core of Precision Farming (Geng & Dong, 2017; Pathan et al., 2020). Soil quality monitoring is imperative for its health and potential imbalances that can affect the crop. Soil moisture is another significant parameter in agricultural operations which is crucial for the management of water resources and drought control. A DL regression network was used to construct a soil moisture prediction model that later can be used to develop effective strategies for water-saving and controlling drought (Cai et al., 2019). Computer vision and DL algorithms are being employed for the processing of data captured by drones and/or software-based technology to monitor soil and crop health. A DL application called Plantix developed by Berlin-based agricultural tech start-up PEAT spots potential defects and nutrient deficiencies in soil associating foliage patterns with plant disease and pest infestation as well as soil deficiencies. California-based Trace Genomics also performs soil analysis and diagnostics through comprehensive microbial evaluation.

37.5.1.9 Nutrient monitoring

Nutrient monitoring, as the name suggests, involves monitoring nutrients, pH, alkalinity in water bodies, soil, hydroponics as well as plants with the aid of wireless sensors and IoT devices by estimating mainly nitrogen, phosphorous, and potassium for increasing crop yield (Dolci, 2017). Nutrient deficiencies can cause moderate to severe crop loss; hence it is necessary to monitor for providing suitable intervention for preventing crop losses. The plant nutrient monitoring is done at three levels, that is, at soil/water level (what is given to the plants), at root zone level (what is available to plants), and a leaf/stem level (what ends up in the plant). Water used for irrigation can be from water bodies, hydroponics is farming without soil; therefore nutrients in the water are significant for successful hydroponic cultivation and higher field yield. NexSens UV nitrate sensors in tandem provide a monitoring system to better understand seasonal nutrient loads in water bodies (Burton et al., 2018). Nutrient Film Technique, one of the hydroponics techniques, uses a nutrient solution to drain on the root area and has been successfully applied in lettuce cultivation. Similarly, nutrients in the soil are also crucial not only for higher crop yield but also for their disease-free or health status. Electrical conductivity, cation, and anion exchange capacities are used for estimating N, P, K, Ca, Mg, and S.

37.5.1.10 Smart greenhouse management

The smart or automated greenhouse management system can monitor climatic conditions and carry out robotic crop treatments by utilizing AI with IoT technologies on a potentially large-scale for cultivation. They are customized and specialized microfarming solution for individual farmers that incorporates sensors and actuators to fully automated greenhouses which will additionally help in protecting against the external and environmental factors influence. Therefore the greenhouses will monitor climatic conditions and carry out robotic treatments accordingly, including soil preparation, sowing, weeding, and crop harvesting and predict from eliminating production errors to optimize production costs for each microfarm (Burton et al., 2018; Yahya, 2018). In 2019, five international horticulture teams joined a smart greenhouse experiment where each team grows a cucumber crop remotely for a 4-month-period in a compartment equipped with standard actuators (ventilation, heating, lighting, screening, fogging, water CO₂, and nutrient supply). Control features were remotely determined by each team using their own AI algorithms, which varied between supervised, unsupervised, and reinforcement ML (Deep Reinforcement Learning, Dynamic Regression, GAN, CNN, RNN). It was concluded that overall AI performed well in controlling a greenhouse and one team even outperformed the manually-grown reference (Hemming et al., 2019).

37.5.1.11 Weather tracking and forecasting

Climate-crop association is very significant in agriculture. The variability in the climate makes accurate forecasting difficult. Currently, ML, DL algorithms, and specific crop models are being used to decode, forecast, and understand data-intensive processes in agricultural operational environments (Dolci, 2017). In 2020, an improved data-driven global weather forecasting framework/model using a deep CNN was developed for forecasting constant and accurate weather patterns of several weeks and longer duration by using atmospheric state variables as input. The model computes realistic forecasts and can be executed quickly; this further offers a potential opportunity for future developments (Weyn et al., 2020).

37.5.2 As a service industry

There's no doubt that technology is becoming faster, smarter, better; but many professions, by their very nature, from artists and writers to doctors and nurses require human intervention; essentially due to the compassion, empathy, trust, and personality which to date cannot be programmed into a machine. Currently, both China and the United States are frontiers in adopting AI, with a 190% increase in the AI patents granted in the last five years and by the investing \$10 billion approx. respectively. Similarly, Russia also aims to make 30 percent of its military equipment robotic by 2025 and the United Kingdom is an investment of over £603 million in the AI industry (Marinchak et al., 2018). Out of all AI fields, the field and customer or advisory service is the one where AI is hugely embraced in the form of chatbots. AI enables "predictive field service," which anticipates service requirements and automatically adjusts processes accordingly. For example, in agriculture, while tasks that are simple and monotonous can be easily automated, AI solutions can be used to enhance the ability to decide based on recorded observations of a variety of parameters.

The AI-as-a-Service (AIaaS) refers to third-party that provide out-of-the-box AI solutions as a cost-effective alternative to developing AI software. AIaaS makes AI technology accessible to everyone even without writing code (Beimbom et al., 2011). The popular solution includes Bots and APIs which utilize ML, NLP, and computer vision.

Chatbots use AI algorithms to simulate human conversation by NLP and ML for understanding user queries and provide relevant responses. AI-powered chatbots assist in a high-quality personalized experience, support, speed and efficiency, and cost-saving. Currently, major industries that rely on AI in customer support or advisory service are food, travel, finance, retail, airline, and clothing (Ivanov & Webster, 2017). Therefore similar can be applied in Agriculture where farmers can get the information regarding best sowing, irrigation, and harvesting time based on the weather forecasting. On a smaller scale, it has been implemented for farmers of few dozen villages in Karnataka, Andhra Pradesh, Telangana, Madhya Pradesh, and Maharashtra through a joint effort Microsoft India and International Crop Research Institute for the Semiarid Tropics (ICRISAT). Technologies like Cloud ML, Satellite Imagery, and Advanced Analytics are optimized by Microsoft Cortana Intelligence Suite for providing higher crop yield and better price control using data from geostationary satellite images without installing sensors implicitly on farmer's field.

37.5.3 Biological sciences

Biological sciences are one of the most promising beneficiaries of AI. Currently, the use of Bioinformatics and AI algorithms like ML, DL, ANN among others in exploring genetic mutations responsible for physiological changes to examining pathological effects. Biology generates immensely large, complex, and convoluted data that offers valuable insights that could be used to improve our health with proper investigation and applications.

37.5.3.1 Bioinformatics

Bioinformatics is a multidisciplinary field to enhance understanding of biological data through developing methods, tools, or software. With the aid of AI, ML, and DL, it can achieve its numerous objective such as gene expression analysis, protein classification, prediction, and pattern detection among others by utilizing varied datasets. Typically, a bioinformatics approach includes predictive analytics where a query is searched against a previously known dataset or annotation for forming any conclusive information. For instance, in protein structure and gene prediction or gene finding predictive models that are generated using ML/DL algorithm such as K-Nearest Neighbor, SVM and neural networks among others methodically searches the genomic DNA for protein-coding genes (Asgari & Mofrad, 2015; Kelley et al., 2012; Senior et al., 2020). Recent advances also employ computer vision and Deep Convolutional Neural Networks for the identification of protein families and their subsequent classification. For instance, the protein pattern classifier model developed by Optima AI derives multilabel multiclass classifications using high throughput cellular microscopy images. Other subfields of bioinformatics where AI can play a significant role are Sequence/ Structure/ Functional analysis, integrative bioinformatics, protein interaction networks, metabolic networks, and pathway analysis. Biomarker discovery, Pharmacogenomics, Functional Enrichment Analysis from omics data are the new emerging disciplines of Bioinformatics with the aid of AI.

37.5.3.2 Molecular biology and omics data mining

The advancement in omics technology has resulted in an expansion of molecular data in modern research. The deviations of normal physiological processes are reflected by genes, DNAs, RNAs, proteins, and other biomolecules expressions and profiles in different types of omics data. All the biological entities, that is, DNAs, RNAs, proteins, and other biomolecules are immensely correlated; hence, the integrative analysis of multiomics data is required for making sense of them. AI algorithms such as ML and DL has the adeptness to make decisive interpretation of this enormously complex data and currently seems like the most effective tool for the analysis and understanding of multiomics data. For instance, the combination of medical images and clinical datasets with the omics data has promoted precision and personalized medicine by creating predictive models and identifying patterns using computer vision and DL (Ahmed, 2020; Martorell-Marugán et al., 2019). In recent years the ML and DL are popularly being used in cancer and muscle disease diagnostics and drug discovery (Preusse et al., 2020). Individumed, founded in Hamburg, Germany in 2002, is an integrated AI Platform that includes tools for immuno-oncology, clinical data, genomics, expression, and pathway analytics. PHARMA.AI, a fully AI-integrated drug discovery software suite that can be used for disease target identification, synthetic biology generation and generation of novel molecules data, and predicting clinical trial outcomes. Methods of identifying cancer progression is also being developed where epigenetic markers, that is, promoter methylation is established using epigenomics studies and ML for recognizing transcriptional accessibility and molecular processes involved in development, tissue maintenance, disease states, and eventually aging (Chen et al., 2019).

Currently, the use of AI in omics is limited to phenomics which deals with phenome, that is, the entire phenotypes expressed by a cell, tissue, organ, organism, or species primarily by using algorithms of the computer vision, DL and

ML. Deep Plant Phenomics is a DL Platform developed for complex plant phenotyping tasks including leaf counting for establishing the baseline for the mutant classification and age regression. Verily Life Sciences (formerly Google Life Sciences and a subsidiary of Google's parent company, Alphabet) in San Francisco, has developed a DL tool "DeepVariant" for identification of SNPs from low-quality reference genomes and achieved an error rate of only $\sim 2\%$ as compared to 20% through conventional methods (Zhao et al., 2020). Atomwise, another San Francisco-based company, performs AI-driven molecular-screening using DL algorithms by converting the 3D structure of proteins and small molecules into 3D pixel grids leading to high atomic and geometries precision. This further aids in predicting the likely interaction between small molecules and a given protein. DL algorithms unravel categorization challenges by examining molecular features like shape and hydrogen bonding among others for recognizing key criteria and ranking potential drugs. Deep Genomics, Toronto-based company uses genomics and transcriptomics data from healthy cells for predicting disease progression and treatment. DL algorithms based predictive models of RNA-processing events like transcription, splicing, and polyadenylation using input datasets and applying on clinical data recognizes changes and flag them as pathogenic (Rothman & Kraft, 2006). Even though the use of AI in plant omics is in the nascent stage, AI can also be used in plant omics for recognizing the pattern, biomolecules, and molecular processes for a better understanding of biological processes.

37.5.3.3 System and synthetic biology

Recently, AI is introduced in the field of system and synthetic biology which deals with understanding at the organism, tissue, or cell level by putting all of their pieces together and engineering biological systems, respectively. The AI can help by designing more effective experiments and decisive data analytics for identifying novel genetic circuits under system biology and engineering living organisms with new functions under synthetic biology. One major bottleneck of biological experiments is the reductionist approach used in the majority of research experiments, that is, target generally is a piece of biological process and is very specific; this causes a hinge during data analysis. The need for DL and ML-based algorithms in biology is to both experiment designing and data analysis of varied data types (Artificial Intelligence Methods & Tools For Systems Biology, 2004; Nesbeth et al., 2016). Using Riffyn's ML and cloud-based software platform, the designing, and standardization of experiments along with data analysis (or de novo) that are significant as industrial enzymes to crops and their microbiomes utilizing its ML/DL based protein design platform. DL techniques ensure correct folding and function of designed proteins. On completion of engineering, the new proteins are produced by fermentation, bypassing the natural evolution of producing brand-new molecules. Distributed Bio, founded in 2012 and based in San Francisco, utilizes the Tumbler platform through ML methods for revolutionizes therapeutics by engineering protein to optimize existing antibodies. It created more than 500 million antibody variants, scores them as per their binding to potential targets, and identify valuable changes for further improvement of antibodies, and synthesis as well as testing of the top scorers (Artificial Intelligence Methods & Tools For Systems Biology, 2004; Nesbeth et al., 2016). Synthetic biology has the potential to make noteworthy effects in virtually every sector: food, agriculture, medicine, climate, energy, and materials.

37.6 Future perspective and challenges

The application of AI in several industries such as technology, banking, marketing, entertainment, and little in agriculture has seen success. Even with many success stories, AI is affected by numerous issues and challenges in AI implementation such as limitation in computing power, trust deficit, knowledge, data privacy and security, algorithms bias, and data scarcity. The amount of power required for implementing AI-related projects is one of the keys that is keeping most developers away. Both ML and DL have an unlimited demand for the number of cores and GPUs to work efficiently. Even with the cloud computing and massively parallel processing systems, which did provide hope for increased AI implementation but as the data volumes go up, and DL moves toward automated creation of increasingly complex algorithms, cloud computing will also reach its limits. The expenses incurred for AI computing are not easily affordable and are bound to shore up as the inflow of unprecedented amounts of input data owing to the rapidly increasing complex algorithms.

One of the most important factors that are a cause of worry for AI is the unknown nature of how DL models predict the output which results in the lack of trust for AI solutions. Another factor that presents a significant challenge is the goal of achieving human-like through AI. Every learning model enjoying $\sim 90\%$ accuracy is easily lost to humans. For instance, the model predicting whether an image is of a cat or a dog. Humans can easily achieve the spectacular $>99\%$ accuracy; but to achieve the same through ML or DL would require extreme optimization, finetuning, a large dataset,

and a well-defined and precise algorithm, along with exceptional computing power, uninterrupted training on train data, and testing on test data. But still, even the specific pretrained models, trained on millions of images and are fine-tuned for maximum accuracy, continue to show errors and struggle to grasp human-level performance.

Every AI model is based on the data collected from users around the world. Most of the AI applications are based on massive volumes of data to learn and make intelligent decisions. Often the ML systems depend on sensitive and personal data for learning and improving themselves. Due to this systematic learning process, these ML systems can become prone to data breaches and identity theft. The data collected can be used for good or bad purposes. As the amount of data increases the storage and security requirement is also growing. To address this some companies have started to train data on the smart device and only trained model is sent to servers, not the data itself. Still, not every AI implementer can afford to go this route. The good or bad aspect of an AI system is directly depending on the amount of data they are trained on; good data will lead to a good AI solution whereas the bad will lead to a bad solution. Everyday data collected for AI training is generally of bad quality and holds no significance of its own. As the input data of poor quality therefore the resulting solution will also reflect that quality and bias. For instance, if bad data is associated with, communal, ethnic, gender, or racial biases is used for training the AI interference model then the model will also carry that bias might lead to unethical and unfair results. Therefore it is emergent that the use for training AI models is unbiased.

Data scarcity is primarily due to the stringent IT rules established by governments around the globe; as major enterprises are facing charges for unethical use of user data generated. The data is a core of AI, and labeled data is used for training, learning, and making predictions. A few companies are innovating new methodologies that can give precise results despite the data scarcity. But it is well known that biased information can only lead to a biased and flawed AI system.

Regardless of the type of AI, or its application, it is in the midst of a computational revolution and soon its tendrils will be beyond the “computer world.” In the coming decades, the application of AI will increase manifold with its fast adoption rates. For that to achieve hardware with greater performance, computational power, cost efficiency for training sophisticated are need to be developed. Besides silicon chips, other materials and newer architecture for cloud and edge computing are the focus of current AI research.

37.7 Conclusion

The impact of AI on human lives and the economy has been astonishing. AI contribution toward the world economy is projected to be around ~\$15.7 trillion by 2030. To take that into perspective, that’s about the combined economic output of China and India as of today. With various companies predicting that the use of AI can boost business productivity by up to 40%, the dramatic increase in the number of AI start-ups has magnified 14 times since 2000. The AI application and potential are widespread from automation to complex prediction to agriculture to health much more. Although AI challenges seem very depressing and devastating for mankind, through the collective effort of people, they can be effectively tackled.

References

- Abadi M. (2016). TensorFlow: Learning functions at scale. In *Proceedings of the twenty-first ACM SIGPLAN international conference on functional programming*. (p. 1).
- Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., et al. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the twelfth USENIX symposium on operating systems design and implementation, OSDI 2016*. (pp. 265–283).
- Aggarwal M., Murty M.N. (2021). Deep learning. In: *Springerbriefs in applied sciences and technology*, (pp. 35–66).
- Ahmad J., Farman H., Jan Z. (2019). Deep learning methods and applications. In: *Springerbriefs in computer science*. (pp. 31–42).
- Ahmed, Z. (2020). Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. *Human Genomics*, 14(1).
- Ambrogio S., Narayanan P., Tsai H., MacKin C., Spoon K., Chen A., et al. (2020). Inference of deep neural networks with analog memory devices. In *Proceedings of the international symposium on VLSI technology, systems and applications, VLSI-TSA 2020*, (pp. 119–120).
- Dubitzky W., Azuaje F. (2004). *Artificial intelligence methods and tools for systems biology*, Springer.
- Asgari, E., & Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*, 10(11).
- Bakshi S., Johnsson L. (2020). A highly efficient SGEMM implementation using DMA on the intel/movidius myriad-2. In *Proceedings of the symposium on computer architecture and high performance computing*, (pp. 321–328).
- Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, 43(1), 3–31.

- Bavikadi S., Sutradhar P.R., Khasawneh K.N., Ganguly A., Dinakarrao S.M.P. (2020). A review of in-memory computing architectures for machine learning applications. In *Proceedings of the ACM Great Lakes symposium on VLSI, GLSVLSI*, (pp. 89–94).
- Beimborn, D., Miletzki, T., & Wenzel, S. (2011). Platform as a service (PaaS). *Wirtschaftsinformatik*, 53(6), 371–375.
- Bhatnagar S., Prasad H., Prashanth L. (2013). Reinforcement learning. In: *Lecture notes in control and information sciences*, (pp. 187–220).
- Böhm C., Noll R., Plant C., Zherdin A. (2009). Index-supported similarity join on Graphics processors. In: *Datenbanksysteme in business, technologie und web, BTW 2009 - thirteenth Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, proceedings, (pp. 57–66).
- Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1–2), 119–155.
- Brewka, G. (1996). Artificial intelligence—A modern approach by Stuart Russell and Peter Norvig, Prentice Hall, Series in Artificial Intelligence, Englewood Cliffs, NJ In (third ed.S. Russel, & P. Norvig (Eds.), *The Knowledge Engineering Review* (Vol. 11 Pearson Education limited.
- Bundy, A. (2017). Preparing for the future of artificial intelligence. *Artificial Intelligence SoC*, 32(2), 285–287.
- Burton, L., Dave, N., Fernandez, R. E., Jayachandran, K., & Bhansali, S. (2018). Smart gardening IoT soil sheets for real-time nutrient analysis. *Journal of the Electrochemical Society*, 165(8), B3157–B3162.
- Buscema P.M., Massini G., Breda M., Lodwick W.A., Newman F., Asadi-Zeydabadi M. (2018). Artificial neural networks. In *Studies in systems, decision and control*, (pp. 11–35).
- Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., & Xue, X. (2019). Research on soil moisture prediction model based on deep learning. *PLoS One*, 14(4).
- Chang, C. W., Lee, H. W., & Liu, C. H. (2018). A review of artificial intelligence algorithms used for smart machine tools. *Inventions*, 14.
- Chen, A. (2016a). A review of emerging non-volatile memory (NVM) technologies and applications. *Solid State Electron*, 125, 25–38.
- Chen, J. X. (2016b). The evolution of computing: AlphaGo. *Computing in Science and Engineering*, 4–7.
- Chen, S., Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12), 1452–1457.
- Cholaquidis, A., Fraiman, R., & Sued, M. (2020). On semi-supervised learning. *Test*, 29(4), 914–937.
- Coates A., Baumstarck P., Le Q., Ng A.Y. (2009). Scalable learning for object detection with GPU hardware. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, IROS 2009*, (pp. 4287–4293).
- Cruz, A., Ampatzidis, Y., Pierro, R., Materazzi, A., Panattoni, A., De Bellis, L., et al. (2019). Detection of grapevine yellows symptoms in *Vitis vinifera* L. with artificial intelligence. *Computers and Electronics in Agriculture*, 157, 63–76.
- Cunningham P., Cord M., Delany S.J. (2008). Supervised learning. In *Cognitive technologies*, (pp. 21–49).
- Davenport T.H., Ronanki R. (2018). Artificial intelligence for the real world. In *Harvard business review*.
- Devroye L., Lugosi G. (2001). *Minimax theory*, (pp. 150–176).
- Ding Z., Huang Y., Yuan H., Dong H. (2020). Introduction to reinforcement learning. In *Deep reinforcement learning: Fundamentals, research and applications*, (pp. 47–123).
- Dligach, D., Miller, T., & Savova, G. K. (2015). Semi-supervised learning for phenotyping tasks. *AMIA. Annual Symposium Proceedings, 2015*, 502–511.
- Does. (2018). AI have a hardware problem? *Nature Electronics*, 205.
- Dolci R. (2017). IoT solutions for precision farming and food manufacturing: Artificial intelligence applications in digital food. In *Proceedings of the international computer software and applications conference*, (pp. 384–385).
- Elfving, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107, 3–11.
- Francis L. (2014). Unsupervised learning. In *Predictive modeling applications in actuarial science: Volume I: Predictive modeling techniques*, (pp. 280–312).
- Fritz, S., See, L., McCallum, I., You, L., Bun, A., Moltchanova, E., et al. (2015). Mapping global cropland and field size. *Global Change Biology*, 21(5), 1980–1992.
- Gauchi R., Kooli M., Vivet P., Noel J.P., Beigne E., Mitra S., et al. (2019). Memory sizing of a scalable SRAM in-memory computing tile based architecture. In *Proceedings of the IEEE/IFIP international conference on VLSI and system-on-chip, VLSI-SoC*, (pp. 166–171).
- Geng, L., & Dong, T. (2017). An agricultural monitoring system based on wireless sensor and depth learning algorithm. *International Journal of Online and Biomedical Engineering*, 13(12), 127–137.
- Goldberg, X. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6, 1–116.
- Ha Y.K., Lee J.E., Kim H.J., Bae J.S., Oh S.C., Nam K.T., et al. (2004). MRAM with novel shaped cell using synthetic anti-ferromagnetic free layer. In *Digest of technical papers - Symposium on VLSI technology*, (pp. 24–25).
- Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61(4), 5–14.
- Hansson A., Agarwal N., Kolli A., Wenisch T., Udipi A.N. (2014). Simulating DRAM controllers for future system architecture exploration. In *Proceedings of the ISPASS - IEEE international symposium on performance analysis of systems and software*, (pp. 201–210).
- Hemming, S., De Zwart, F., Elings, A., Righini, I., & Petropoulou, A. (2019). Remote control of greenhouse vegetable production with artificial intelligence—Greenhouse climate, irrigation, and crop production. *Sensors (Switzerland)* (8), 19.
- Hickmann B., Chen J., Rotzin M., Yang A., Urbanski M., Avancha S. (2020). Intel Nervana neural network processor-T (NNP-T) fused floating point many-term dot product. In *Proceedings of the symposium on computer arithmetic*, (pp. 133–136).
- Hlavac M., Maymin S., Breazeal C. (2004). *Artificial intelligence platform*. Google Patents. 696 p.
- Ivanov S., Webster C. (2017). Adoption of robots, artificial intelligence and service automation. In *Proceedings of the international scientific conference. CONTEMPORARY TOURISM – TRADITIONS AND INNOVATIONS 19- 21 Oct 2017, Sofia Univ.* (pp. 1–9).

- Jaakkola H., Henno J., Mäkelä J., Thalheim B. (2019). Artificial intelligence yesterday, today and tomorrow. In *Proceedings of the fourth-second international convention on information and communication technology, electronics and microelectronics, MIPRO 2019*, (pp. 860–867).
- Jia Z., Tillman B., Maggioni M., Scarpazza D.P. (2019). *Dissecting the graphcore IPU architecture via microbenchmarking*. arXiv.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)*, 255–260.
- Jun H., Cho J., Lee K., Son H.Y., Kim K., Jin H., et al. (2017). HBM (High bandwidth memory) DRAM technology and architecture. In *Proceedings of the IEEE ninth international memory workshop, IMW 2017*.
- Kacher I., Portaz M., Randrianarivo H., Peyronnet S. (2020). *Graphcore C2 card performance for image-based deep learning application: A report*. arXiv.
- Kalischuk, M., Paret, M. L., Freeman, J. H., Raj, D., Da Silva, S., Eubanks, S., et al. (2019). An improved crop scouting technique incorporating unmanned aerial vehicle-assisted multispectral crop imaging into conventional scouting practice for gummy stem blight in Watermelon. *Plant Disease*, 103(7), 1642–1650.
- Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., & Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research*, 40(1).
- Ketkar N., Ketkar N. (2017). Introduction to tensorflow. In *Deep learning with Python*, (pp. 159–194).
- Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10.
- Khatab, A., Habib, S. E. D., Ismail, H., Zayan, S., Fahmy, Y., & Khairy, M. M. (2019). An IoT-based cognitive monitoring system for early plant disease forecast. *Computers and Electronics in Agriculture*, 166.
- Kim, C., Lee, H. W., & Song, J. (2016). Memory interfaces: Past, present, and future. *IEEE Solid-State Circuits Magazine*, 8(2), 23–34.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana)*, 249–268.
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 436–444.
- Lee, V. W., Kim, C., Chhugani, J., Deisher, M., Kim, D., Nguyen, A. D., et al. (2010). Debunking the 100X GPU vs. CPU myth. *ACM SIGARCH Computer Architecture News.*, 38(3), 451–460.
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors (Switzerland)*.
- Liao, C., Wang, J., Dong, T., Shang, J., Liu, J., & Song, Y. (2019). Using spatio-temporal fusion of Landsat-8 and MODIS data to derive phenology, biomass and yield estimates for corn and soybean. *The Science of the Total Environment*, 650, 1707–1721.
- Littman M.L. (2015). Markov decision processes. In *International encyclopedia of the social & behavioral sciences*, second edition. (pp. 573–575).
- Lloyd Y. (2015). *Valeo and Mobileye sign a unique technology cooperation agreement*. Press release, 3.
- Lorenc, R. S., Beskid, M., Poniatowski, L. W., & Jerzmanowska, M. (1973). The use of silica gel for human calcitonin isolation in some thyroid diseases. *Endocrinologia Experimentalis*, 267–273.
- Magalhães, P. S. G., & Cerri, D. G. P. (2007). Yield monitoring of sugar cane. *Biosystems Engineering*, 96(1), 1–6.
- Malhotra P., Vig L., Shroff G., Agarwal P. (2015). Long short term memory networks for anomaly detection in time series. In *Proceeding of the twenty-third European symposium on artificial neural networks, computational intelligence and machine learning, ESANN 2015*, (pp. 89–94).
- Marinchak, C. M. D., Forrest, E., & Hoanca, B. (2018). Artificial intelligence: Redefining marketing management and the customer experience. *International Journal of E-Entrepreneurship and Innovation*, 8(2), 14–24.
- Martorell-Marugán, J., Tabik, S., Benhammou, Y., del Val, C., Zwir, I., Herrera, F., et al. (2019). Deep learning in omics data analysis and precision medicine. *Computational Biology*.
- Millstein, R. (1968). The logic theorist in LISP. *International Journal of Computer Mathematics*, 2(1–4), 111–122.
- Mittal, S. (2018). A survey of ReRAM-based architectures for processing-in-memory and neural networks. *Machine Learning and Knowledge Extraction*, 1(1), 75–114.
- Mittal, S. (2020). A survey of FPGA-based accelerators for convolutional neural networks. *Neural Computing and Applications*, 1109–1139.
- Mkonyi, L., Rubanga, D., Richard, M., Zekeya, N., Sawahiko, S., Maiseli, B., et al. (2020). Early identification of Tuta absoluta in tomato plants using deep learning. *Sci African*, 10.
- Mohammed M., Khan M.B., Bashie E.B.M. (2016). *Machine learning: Algorithms and applications*, (pp. 1–204).
- Momose, H., Kaneko, T., & Asai, T. (2020). Systems and circuits for AI chips and their trends. *Japanese Journal of Applied Physics*.
- Moutinho L., Hutcheson G., Lin F.-J. (2014). Clustering algorithms. In *The SAGE dictionary of quantitative management research*. (pp. 38–38).
- Muni, A., & Hansen, J. (2005). Amazon web services. *Dr. Dobb's Journal.*, 66–67.
- Murase H. (2000). Artificial intelligence in agriculture. In *Computers and electronics in agriculture*. (pp. 1–2).
- Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief history of artificial intelligence. *Neuroimaging Clinics of North America*, 393–399.
- Nelli F., Nelli F. (2018). Deep learning with TensorFlow. In: *Python data analytics*, (pp. 349–407).
- Nesbeth, D. N., Zaikin, A., Saka, Y., Romano, M. C., Giuraniuc, C. V., Kanakov, O., et al. (2016). Synthetic biology routes to bio-artificial intelligence. *Essays in Biochemistry*, 60(4), 381–391.
- O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE Intelligent Systems*, 28(2), 96–99.
- Okwu M.O., Tartibu L.K. (2021). Artificial neural network. In *Studies in computational intelligence*, (pp. 133–145).
- Panesar A., Panesar A. (2019). What is machine learning? In *Machine Learning and AI for Healthcare*, (pp. 75–118).
- Pantazi X.E., Moshou D., Bochtis D. (2020). Artificial intelligence in agriculture. In *Intelligent data mining and fusion systems in agriculture*, (pp. 17–101).

- Pathan, M., Patel, N., Yagnik, H., & Shah, M. (2020). Artificial cognition for applications in smart agriculture: A comprehensive review. *Artificial Intelligence in Agriculture*, 4, 81–95.
- Preusse, C., Ross, A., Hathazi, D., Hentschel, A., Goebel, H., & Stenzel, W. (2020). OMICs and AI approaches for muscle diseases. *Neuromuscular Disorders: NMD*, 30, S48.
- Quigley M., Berger E., Ng A.Y. (2007). STAIR: Hardware and software architecture. In: *AAAI workshop—Technical report*. (pp. 31–37).
- Radhakrishnan S., Kolippakkam D., Mathura V.S. (2007). Introduction to algorithms. In: *Bioinformatics: A concept-based introduction*, (pp. 27–37).
- Roosjen, P. P. J., Kellenberger, B., Kooistra, L., Green, D. R., & Fahrretrapp, J. (2020). Deep learning for automated detection of *Drosophila suzukii*: potential for UAV-based monitoring. *Pest Management Science*, 76(9), 2994–3002.
- Rothman, H., & Kraft, A. (2006). Downstream and into deep biology: Evolving business models in “top tier” genomics companies. *Journal of Commercial Biotechnology*, 12(2), 86–98.
- Sajja P.S. (2021). Introduction to artificial intelligence. In *Studies in computational intelligence*, (p. 1–25).
- Schapire, R. E. (2013). Boosting: Foundations and algorithms. *Kybernetes*, 164–166.
- Schmid, F., Frommberger, L., Cai, C., & Freksa, C. (2013). What you see is what you map: Geometry-preserving micro-mapping for smaller geographic objects with MAPIT. *Lecture Notes in Geoinformation and Cartography*, 3–19.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 85–117.
- Schwichtenberg, H., & Schwichtenberg, H. (2020). Microsoft Azure. In: *Windows PowerShell 5 und PowerShell*, 7, 1155–1202.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710.
- Seyedzadeh, A., Maroufpoor, S., Maroufpoor, E., Shiri, J., Bozorg-Haddad, O., & Gavazi, F. (2020). Artificial intelligence approach to estimate discharge of drip tape irrigation based on temperature and pressure. *Agricultural Water Management*, 228.
- Shaw G.L. (1986). Donald Hebb: The organization of behavior. In: *Brain theory*, (pp. 231–233).
- Shekhar, Y., Dagur, E., Mishra, S., Tom, R. J., Veeramankandan, M., & Sankaranarayanan, S. (2017). Intelligent IoT based automated irrigation system. *International Journal of Applied Engineering Research*, 12(18), 7306–7320.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359.
- Stalzer M.A. (2012). Flashblades: System architecture and applications. In *ACM international conference proceeding series*, (pp. 10–14).
- Sundvall S. (2019). Artificial intelligence. In *Critical terms in futures studies*, (pp. 29–34).
- Szepesvári C. (2010). Algorithms for reinforcement learning. In *Synthesis lectures on artificial intelligence and machine learning*, (pp. 1–89).
- Tarafdar M., Beath C.M. (2018). Wipro limited: Developing a cognitive DNA. In *Proceedings of the international conference on information systems, ICIS 2018*.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 44–56.
- Tran D., Bourdev L., Fergus R., Torresani L., Paluri M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, (pp. 4489–4497).
- Tsai, W. T., Bai, X. Y., & Huang, Y. (2014). Software-as-a-service (SaaS): Perspectives and challenges. *Science China Information Sciences*, 57(5), 1–15.
- Tsymbal E.Y., Žutić I., Åkerman J. (2019). Magnetoresistive random access memory. In: *Spintronics handbook: Spin transport and magnetism*, second ed., (pp. 421–442).
- Tu Y. (2019). Machine learning. In: *EEG signal processing and feature extraction*, (pp. 301–323).
- Turing A.M. (2012). Computing machinery and intelligence. In: *Machine intelligence: Perspectives on the computational model*, (p. 1–28).
- Upadhyay, N. K., Jiang, H., Wang, Z., Asapu, S., Xia, Q., & Joshua Yang, J. (2019). Emerging memory devices for neuromorphic computing. *Advanced Materials Technologies*.
- Varghese, A., Edwards, B., Mitra, G., & Rendell, A. P. (2017). Programming the adapteva Epiphany 64-core network-on-chip coprocessor. *International Journal of High Performance Computing Applications*, 31(4), 285–302.
- Varia, J., & Mathew, S. (2014). Overview of Amazon Web Services (Survey Report). *Seminar Nasional Aplikasi Teknologi Informatika, 2010*(January), 1–30.
- Vieira S., Lopez Pinaya W.H., Mechelli A. (2019). Introduction to machine learning. In: *Machine learning: Methods and applications to brain disorders*, (pp. 1–20).
- Welsler J., Pitera J.W., Goldberg C. (2019). Future computing hardware for AI. In *Technical digest - International electron devices meeting, IEDM*. (pp. 1.3.1-1.3.6).
- Weyn, J. A., Durrant, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9).
- Wilamowski B.M., Irwin J.D. (2016). *Intelligent systems*, (pp. 1–596).
- Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165–193.
- Xue C.J., Zhang Y., Chen Y., Sun G., Yang J.J., Li H. (2011). Emerging non-volatile memories: Opportunities and challenges. In: *Embedded systems week 2011, ESWEK 2011 - Proceedings of the ninth IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis, CODES + ISSS'11*, (pp. 325–334).
- Yahya N. (2018). Agricultural 4.0: Its implementation toward future sustainability. In: *Green energy and technology*, (pp. 125–145).

- Yang Z., Moczulski M., Denil M., De Freitas N., Smola A., Song L., et al. (2015). Deep fried convnets. In: *Proceedings of the IEEE international conference on computer vision*, (pp. 1476–1483).
- Yao, X. (1999). Evolving artificial neural networks. *Proc IEEE*, 87(9), 1423–1447.
- Zhao B. (2020). The application of artificial intelligence in agriculture. *Journal of Physics: Conference Series*.
- Zhao S., Agafonov O., Azab A., Stokowy T., Hovig E. (2020). *Accuracy and efficiency of germline variant calling pipelines for human genome data*. *bioRxiv*.
- Zheng, A. (2015). *Evaluating machine learning algorithms. A beginner's guide to key concepts and pitfalls*, (p. 59) Springer.
- Zhou D.-X. (2015). Machine learning algorithms. In: *Encyclopedia of applied and computational mathematics*, (pp. 839–841).

Index

Note: Page numbers followed by “f” and “t” refer to figures and tables, respectively.

A

ABA biosynthesis-related genes, 368
Abiotic factor, 441
Abiotic stress, 5, 126–127, 126f, 163, 270–272
 expression quantitative trait loci (eQTL), 561
 G-protein signaling, 561
 metabolomics, 126–127, 126f
 Na⁺ transport and extrusion, 561
 potato (*Solanum tuberosum* L.), 347, 350
 quantitative trait loci (QTLs), 562
 RdDM pathway, 561–562
 retrograde signaling, 561
 RNA-Seq, 163
 salinity, 561
 tomato, 270–272
Abnormal plant size, 424
Accessible chromatin regions (ACRs), 571–572
Accuracy, of system, 623
Acetylation, 115–116
Active transmission of Apple scar skin viroid (ASSVd), 373
Advanced sensing technologies, 625–626
Affymetrix Axiom framework, 330
Affymetrix GeneChip Rice Genome Array, 45
Agricultural robotics, 625–626
Airborne platform systems, 630
Akida neural processor, 653
Alignment-free methods, 305
Alphaproteobacterium, 407
Alternanthera mosaic virus (AltMV), 426–427
Alternative genome scale approaches, 210–212
Alternative 3' splice site (A3SS), 368
Alternative 5' splice site (A5SS), 368
Alternative splicing (AS), 368
Amazon S3 storage services, 32–33
Amplified fragment length polymorphism (AFLP), 459–460, 463–464, 513
 advantages of, 542–543
 application of, 544
 disadvantages of, 543
 linkage mapping, 544
 phylogenetic methods, 544
 population-based methods, 544
Analysis and GenBank Submission, 423
ANOVA (analysis of variance), 563
Anthesis-silking interval (ASI), 279
Anthocyanins, 266
Antiferromagnetic (AFM) materials, 654

Applied biosystems (ABI) SOLiD sequencer, 445
Arabidopsis, 192
Arabidopsis Genome Initiative, 47, 50
Arabidopsis thaliana, 37, 297–298
Array tape, 468
Arthrobacter, 409
Artificial General Intelligence (AGI), 643–645
Artificial intelligence (AI)
 application of
 agriculture robot and drones, 660
 crop phenotyping, 660
 crop scouting, 660
 disease detection and diagnosis, 660
 field mapping, 658
 irrigation management, 659
 nutrient monitoring, 661
 smart greenhouse management, 661
 soil management, 660
 weather tracking and forecasting, 661
 yield monitoring, 658–659
 biological sciences, 662–663
 future perspective and challenges, 663–664
 history of, 642–643
 methods and approaches in, 643–650
 artificial neural network, 649
 deep learning, 649–650
 machine learning, 645–649
 service industry, 661–662
 tangible and intangible systems, 641
 technological advancements in
 hardware, 651–654
 software, 655–656
 Turing Test, 641
Artificial Narrow Intelligence (ANI), 643–645
Artificial Neural Networks (ANNs), 591
Artificial Super Intelligence (ASI), 643–645
Association mapping (AM)
 in breeding program, 552
 candidate-gene-based, 551–552
 genome-wide association study, 552
 linkage disequilibrium, 551
 methods of, 551, 551f
Astragaloside, 265–266
Atomic absorption spectrometry (AAS), 15
Atomic attractive reverberation (NMR), 491–492
Atrazine, 409
Avsunviroidae, 373

B

Bacterial artificial chromosome (BAC) arrays, 296
Bambino, 43
Barley stripe mosaic virus (BSMV), 390–392, 426–427
Beta-galactosidase, 191–192
Beta-galactosidase (GH35) amino-acid sequences, 199f
β-Galactosidase activity, 191–192
Biallelic mapping, 514
Bidirectional Recurrent Neural Networks (BRNN), 594
Big data, 237
 agriculture ecosystem, 635–636
 in biology and omics, 237
 cloud data source, 631
 internet of things database source, 631
 media source, 631
 remote sensing, 631
 and smart agriculture
 adaptation to climate change, 628
 automated irrigation system, 628
 digital soil and crop mapping, 627
 disease detection and pest management, 628
 fertilizers recommendation, 628
 weather prediction, 627
 sources of, 628–631
 airborne platform systems, 630
 ground platform systems, 630
 remote sensing, 629–630
 statistical data, 630–631
 techniques and tool usage
 cloud platforms, 633
 geographic information systems, 633
 machine learning, 632–633
 vegetation indices, 633–635
 variety, 627
 velocity, 626
 veracity, 627
 volume, 626
Biochemical markers, 538
Biofortification, 298–299
Bioinformatics, 374, 379–380
 brown planthopper resistance in rice, 259
 high-throughput technologies, 37
 human genes, 37
 for microarray data analysis
 Affymetrix Expression Console software, 41

Bioinformatics (*Continued*)

- GeneChip operating software, 41
- microbial pathogens, 37
- next-generation sequencing, 234
- resources for cotton-omics, 237–243, 238*t*
- software, 385–386
- tools, 392–393
- Biomass reaction, 407
- Biopesticides, 221
- Bioremediation, 409
- Biostimulation, 409
- Biotic factor, 441
- Biotic stress, 126–127, 126*f*
- Biparental analysis, 564
- Bisulfite conversion and sequencing (BS-seq), 570
- Bisulfite sequencing, 568
- Boron, 573
- Brachypodium*, 296–297
- Brassica BGALs (BcBGALs), 192
- Breeding, 269
- Brown planthopper (BPH), 253

C

- Caenorhabditis elegans*, 37
- Catechins, 361
- Cauliflower mosaic virus (CaMV), 426–427
- cDNA-amplified fragment length polymorphism (cDNA-AFLP), 256–257
- cDNA synthesis, 365
- CGAP (Chloroplast Genome Analysis Platform), 423
- Chickpea (*Cicer arietinum* L.), 508
- Chilling injury (CI), 130
- Chimeric RNA, 185–186
- China, 361
- ChIP-seq data, 571–572
- Chlorophyll biosynthetic gene (CHLI), 419–420
- Chloroplast photosynthesis related proteins (CPRPs), 428
- Chloroplast protein phosphoglycerate kinase (ChIPGK), 426
- Chloroplasts
 - bioinformatic approaches and plastomes, 422–423
 - chloroplast genome sequencing in plants, 423
 - CMV Y-Sat virus, 419–420
 - colossal damage, 419–420
 - cultivated crops, 419
 - genomic advances, 422
 - nitrogen and sulfur, 419
 - photosynthesis and biosynthesis, 419
 - plant pathogenic viruses, 428–429
 - plant–virus life cycle, 426–428
 - plant–virus metagenomics, 429–430
 - reduced chlorophyll pigmentation, 419–420
 - structural and functional changes, 427–428
 - structure and gene content, 420–421
 - tobacco nuclear genome, 419–420
 - viral infection, 419–420, 427
 - viral infection symptoms, 423–426
- Chromatin immunoprecipitation (ChIP), 571–572
- Chromomethylase 2 (CMT2), 569
- Chromomethylase 3 (CMT3), 569
- Circular RNA, 184–185
- Cis-eQTLs, 565
- Cis-regulatory elements (CREs), 569
- Classical markers
 - biochemical markers, 538
 - cytological markers, 538
 - morphological markers, 538
- CLC Genomics Workbench, 379–380
- Cleaved amplified polymorphic sequence (CAPS), 459–460
- Cleaved amplified polymorphic sequences, 464, 562–563
- Climate change, 107
- Cognitive genomics, 62
- Coleus blumei viroid* (CBVd), 379
- Colloidal CBB, 110
- Combination of restriction enzymes (CORE), 563
- Comparative genomics, 62, 209
- Complexity Reduction of Polymorphic Sequence (CroPS), 563
- Computational biology, 589
- Computational skills, 579
- Computational units, 577
- Computomics
 - applications, 16
 - challenges, 16
- Conservation genomics, 71
- Constraint-based reconstruction and analysis (COBRA), 406–407
- Conventional agriculture, 123
- Conventional data analysis strategies, 593
- Conventional plant breeding approach, 283, 513
- Convolutional Neural Network (CNN), 590
- Coomassie brilliant blue (CBB), 110
- Cotton
 - multiomics data
 - cotton plant diseases, 243–244
 - integration and analysis, 244–245, 244*f*
 - world cotton market, 233
- CpGAVAS (Chloroplast Genome Annotation), 423
- CRISPER-Cas9-based genome-editing, 387
- CRISPER-Cas9 system, 393
- Crop domestication, 209
- Crops production, 129
- Crops quality, 129–130
- C. sinensis* var *sinensis* (CSS), 361–362, 366
- Cucumber disease recognition system, 621–622
- Cucumber mosaic virus (CMV), 419–420
- Culture-independent methods, 441–442
- Culture-independent techniques, 439–440, 441*f*
- Cyanobacteria, 64
- Cytological markers, 538

D

- Data analysis, 169
- Data generation, 165–168
- Data repositories, 245
- dbNSFP tool, 518–520
- dCHIP, 41
- 2D chromatography, 375
- Dead Cas nucleases, 579
- Decision tree regression, 16
- Deep Belief Networks, 594
- Deep learning, 589
 - in agricultural sciences, 600–611
 - in computational biology
 - biological data, in deep neural network, 598–600
 - biological image processing, 595–596
 - multiomic data integration, 596–597
 - omics, 593–594
 - pharmacogenomics, 597–598
 - single-cell RNA sequencing, 597
 - and convolutional neural network, 589–593
- Deep Neural Network (DNN), 598–600, 599*f*
- Degradome, 568
- Dehydration-responsive element-binding (DREB) transcription factor, 573–574
- De novo assembly, 24–25, 49–50, 577
- De novo sequence strategy, 514, 516*f*
 - high-quality raw read, 518, 519*t*
 - input datasets, 517–518
 - quality control, 518
 - Teraclu, 517–518
 - TGICL, 517–518
- Derived cleaved amplified polymorphic sequences, 464
- DeSeq2, 564–565, 567
- Difference-in-gel electrophoresis (DIGE), 107–108
- Differential expression analysis, 170
- Disease resistance, 269–270
- Diversity array technology (DArT Seq), 467, 548, 563
- DNA databases, 44
- DNA decoding techniques, 593
- DnaJ protein, 287–288
- DNA methylation, 569–571
- DNA microarrays, 9
 - applications of, 40
 - bioinformatics tools, 41
 - drawbacks of, 40
- DNA polymorphisms, 6
- DNA sequence homology, 461–462
- DNA sequencing, 6, 61–63, 63*t*, 589
 - high-throughput sequencing, 70
 - shotgun sequencing, 69–70
- DNase-seq, 571–572
- Domain Rearranged Methyltransferase 2 (DMR2), 569
- Doubled haploids, 460
- Double Digested RAD, 563
- Drosophila melanogaster*, 37
- Drought, 279, 280*f*
 - avoidance, 282
 - plant response to, 281–282, 281*f*
 - recovery, 282

resistance in wheat, 331
 stress, 155
 timing, 280–281
 Drought tolerance
 seedling and physiological traits for, 283
 yield traits for, 283

E

Electrospray ionization, 11
 ENCODE, 29
 Epigenomic regulation
 DNA methylation, 569–571
 histone modification, 571–572
 Epigenomics, 23, 29, 62, 64, 240–241
 Epiphany V, 653
 European ash (*Fraxinus excelsior*), 508
 European Bioinformatics Institute (EMBL-EBI), 32–33
 Exon skipping (ES), 368
 Exotic populations, 460
 Expressed sequence tag (EST), 8–9, 38–39, 286, 459–460
 Expression quantitative trait loci (eQTL), 561
 EyeQ, 653

F

FASTA, 379–380
 Flavonoids, 361
 Fluidigm, 467
 Fluorescence-activated nuclei sorting (FANS), 571–572
 Food and Agriculture Organization, 107, 455
 Food security, 107
 Fourier transform, 609
 F₁ population, 460
 F₂ population, 460
 Frequency-based methods, 305
 Fully fermented black tea, 361
 Functional genomics, 62–64, 171, 241–242, 284
 databases, 44–45
 defined, 37–38
 DNA level, 37–38
 dynamic expression of, 493–494
 in fungi, 45
 interactomics, 495
 Malvaceae family plants, 45
 metabolite level, 37–38
 metabolomics, 495
 nutrigenomics, 495–496
 protein level, 37–38
 proteomics, 494–495
 RNA level, 37–38
 transcriptomics, 494

G

Galaxy, 379–380
 Gaussian Mixture Model (GMM), 603
 G + C content analysis, 442
 Gel-based proteomics
 cleaved amplified polymorphic sequence markers, 522, 522*f*, 523*f*

single-stranded conformation polymorphism, 522
 sodium dodecyl sulfate-polyacrylamide gel electrophoresis, 110
 two-dimensional-difference-in-gel electrophoresis, 110–111
 two-dimensional gel electrophoresis, 110
 Gel-based techniques, 107–108
 Gel-free procedures, 107–108
 Gel-free proteomics
 isobaric tags for relative and absolute quantitation, 112
 label-free quantification, 112
 multidimensional protein identification technology, 111
 sequential window acquisition of all theoretical mass spectra, 111
 stable isotope labeling by amino acids in cell culture, 113
 tandem mass tag, 112–113
 GenBank, 44, 379–380
 Gene coexpression network, 361–362
 Gene duplication, 100
 Gene expression regulation, 477–479
 Gene families, 365, 366*t*
 Geneious, 43, 379–380
 Gene mapping, 328
 Gene ontology (GO), 518–520
 Gene predictions, 100
 Generalized Linear models, 564–565
 Gene Regulatory Network (GRN), 569
 Genetic algorithm, 621
 Genetic diversity, 525
 Genetic fingerprinting, 66
 Genetic libraries, 268
 Genetic linkage maps, 66
 Genetic mapping, 525–526, 541
 Genetic maps, 456–458
 Genetic markers for selection, 271
 Genetic transformations, 68
 Gene trait association analysis, 68
 Genome, 61–62
 Genome annotation
 biological databases
 DNA databases, 44
 RNA databases, 44, 44*t*
 functional genomic databases, 44–45
 Genome assembly process, 365*f*
 Genome-by-sequencing (GBS), 3
 Genome coverage, 407
 Genome-editing techniques, 393
 Genome informatics
 in agriculture
 databases and prediction servers, 54–55
 genome assembly, 49–52
 RNA-seq, 52–54
 agrigenomics, 47–48
 computational and statistical techniques, 47
 DNA-seq
 first-generation sequencing technologies, 48
 second-generation sequencing technologies, 48–49
 third-generation sequencing technologies, 49

hit-and-trial experiments, 47
 marker-assisted selection (MAS), 47–48
 structural and functional organization, 47
 Genome mapping
 comparative mapping, 461–462
 molecular marker systems and populations, 459–461
 physical mapping, and genome sequencing, 461
 practical applications of, 462
 Genome-scale metabolic model (GSMM), 406–407
 Genome-wide association (GWA), 300
 Genome-wide association studies (GWAS), 562
 design and analysis, 475–476
 plant and animal breeding, 476
 single-nucleotide polymorphism markers, 474–475
 Genome-wide data analysis, 49–50
 Genome-wide genetic markers, 209
 Genome-wide mutants, 379
 Genomic data, 362*f*
 Genomics, 23, 491, 593
 applications, 9
 applications of, 7–8
 brown planthopper resistance in rice, 253–255, 255*t*, 256*t*
 challenges, 8–10
 cognitive genomics, 62
 comparative genomics, 62
 conservation genomics, 71
 for crop improvement, 71–74, 72*t*
 cyanobacteria, 64
 defined, 61
 DNA sequencing, 61–63, 63*t*
 high-throughput sequencing, 70
 shotgun sequencing, 69–70
 drought stress responses in maize, 284–287
 for enhancing food crops security, 496
 epigenomics, 62, 64, 240–241
 epistasis, pleiotropy, and heterosis, 62*f*
 feature for future breeding, 75
 functional genomics, 62–64, 241–242
 genomic resources
 biparental mapping populations, 65–66
 comparative genome mapping, 66
 functional genomics, 66, 67*t*
 genetic fingerprinting, 66
 genetic linkage maps, 66
 genetic transformations, 68
 gene trait association analysis, 68
 hybrid testing, 66–67
 marker-assisted selection, 68
 molecular markers, 64–65
 transcriptome assemblies, 65, 65*t*
 integrative databases in plants, 69*t*
 in medicine, 71
 metagenomics, 62
 neurogenomics, 62
 NGS technologies for, 6
 nucleomics, 62
 nutrition research, 496–497
 pangenomics, 62, 212
 personal genomics, 62

Genomics (*Continued*)

- plant domestication genomics, 212, 213*t*
- sequence assembly
 - annotation, 71
 - assembly approaches, 70
 - structural genomics, 63
 - synthetic biology and bioengineering, 71
 - transcriptomics, 9, 241
 - translational genomics, 240
 - viruses and bacteriophages, 64
- Genomics-assisted breeding (GAB), 361
- Genomic tools, 210*f*
- Genotype imputation, 331
- Genotyping by sequencing (GBS), 468–469, 563
- GeoChip, 444
- Geographic information systems, 633
- German-based technology, 16
- GeSeq, 423
- GH35, 192
- Glycosylation, 114–115, 197
- GoGene, 518–520
- Grapevine latent viroid (GLVd), 380
- Grapevine yellow speckle viroid*, 378–379
- Graphics, 576
- Green Revolution, 123
- Ground platform systems, 630
- 454 GS 20 Roche sequencing platform, 384
- GWAS, 7

H

- Hardware, AI
 - memory device, 653–654
 - processor, 652–653
 - storage device, 654
- Heat shock proteins (Hsp90), 426
- Helicos BioSciences, 384
- Herbicides, 625–626
 - biodegradation of, 402–409, 404*f*, 406*f*
 - bioremediation of atrazine, 409
 - chemical groups, 399
 - chemical structure, 399
 - 2,4-D (2,4-dichlorophenoxyacetic acid), 400
 - decomposition process, 401
 - degradation, 403*t*
 - environmental impact quotient (EIQ), 399
 - fungi, 401–402
 - glyphosate, 399
 - linuron degradation, 402
 - microbial degradation, 400
 - microorganisms, 401
 - molecular and computational methods, 402
 - phenyl urea herbicides (PUHs), 399
 - photosynthetic system, 399
 - PUHs, 401
 - on targeted plant, 400*f*
 - toxicity of atrazine, 399
 - xenobiotics, 399
- Heterochromatic siRNAs, 568
- HHbits, 379–380
- High affinity K⁺ transporters (HKT), 572–573
- High-fidelity ultradeep sequencing, 376

- High-resolution melting, 465–466
- High-throughput plant phenotyping platforms (HTPPs), 5
- High-throughput posttranslational modification proteomics
 - acetylation, 115–116
 - glycosylation, 114–115
 - phosphorylation, 114
- High-throughput screening, 589
- High-throughput (HTP) fixed single-nucleotide polymorphism microarrays, 467
- High-throughput (HTP) sequence-based markers, 513
- High-throughput sequencing, 47, 378–379, 444, 491–492
- HiSeq 2000, 384
- HiSeq 2500, 384
- HiSeq platform series, 384
- Histone modification, 571–572
- H3K4 methylation, 571
- Hop stunt viroid (HSVd), 376
- Host-selective toxin (HST) functions, 302
- HT sequence-nucleotide polymorphism, 496
- Human Genome Project, 37
- Hybridization-based approaches, 286
- Hybridization based screening, 444
- Hybrid testing, 66–67
- Hymenoscyphus fraxineus*, 508
- Hyperspectral imaging, 620

I

- Illumina paired-end technology, 507
- Image-based plant disease detection, 623
- Image processing, 619
- Industrial agriculture, 123
- Infinium II assay system, 329–330
- Insect pests, 221
- Insect resistance, 270
- Integrated pest management (IPM), 253
- Intel Movidius Myriad, 653
- Intensive agriculture (IA), 124
- Interactomics, 495
- International Human Epigenome Consortium (IHEC), 29
- International Rice Genome Sequencing Project, 47
- International Wheat Genome Sequencing Consortium (IWGSC), 330
 - IWGSC chromosome survey sequence (CSS) assembly, 312
 - IWGSC wheat-genome sequencing project, 312
- Internet of Things (IoT), 602, 625–626, 642
- InterPro, 99
- Intersimple sequence repeat
 - advantages of, 546–547
 - application of, 547
 - disadvantages of, 547
- Inter simple sequence repeat (ISSR), 459–460
- Intron retention, 368
- Ion beam analysis (IBA), 15
- Ionomics, 14–15
- Ion Torrent sequencing, 41

- iPLEX Gold assay, 468
- Iron deficiency, 298
- Isobaric tag-based methodology for relative peptide quantification (iTRAQ), 257–258
- Isobaric tags for relative and absolute quantitation, 112
- I-TASSER, 99

K

- KASP genotyping device, 332–333
- K-means clustering, 609, 622
- K-nearest neighbor (KNN), 16, 609, 621
- Kompetitive allele-specific PCR (KASP) marker technology, 466

L

- Label-free quantification, 112
- Lectin-like SUE domains, 198–200
- LemnaTec Scanalyser, 289
- Linear regression, 16, 563
- Linkage disequilibrium, 551
- Linkage maps, 544, 562–563
- Long noncoding RNA, 181–183, 183*t*
- Long short-term memory (LSTM), 594
- Lutein, 266

M

- Machine learning, 9–10, 236, 589, 632–633
 - in omics
 - genomic studies, 32, 33*f*
 - reinforcement learning, 647–648
 - semisupervised learning, 648–649
 - supervised learning, 646–647
 - unsupervised learning, 647
- Macronutrient, 127
- Magic Viewer, 43
- Magnesium, 128
- Magnetoresistive random-access memory (MRAM), 654
- Maize
 - characteristics of, 279
 - conventional breeding strategies for, 282–283
 - drought stress, 279, 284–290
 - in India, 280
 - in U.S., 279
- Malvaceae family plants, 45
- Map-based cloning, 461
- Marker Assisted Back-Crossing (MABC), 563
- Marker-assisted plant breeding, 513
- Marker-assisted selection (MAS), 47–48, 461, 526–528
 - application of, 552
- Markers
 - biochemical markers, 504, 537–538
 - breeding program, 537
 - cytological markers, 504, 537–538
 - DNA sequence-based markers, 506
 - genetic markers, 506–507
 - biochemical markers, 538
 - classical markers, 538

- molecular markers, 538
 - molecular markers, 537
 - morphological markers, 537
 - polymerase chain reaction (PCR), 537
 - Massively Parallel Signature Sequencing (MPSS) Lynx Therapeutics, 384
 - Mass spectrometry (MS), 43, 235, 257–258
 - Mathematical approach, 30
 - Maxam–Gilbert method, 47, 375
 - Medical Subject Headings (MeSH) vocabularies, 518–520
 - MEME online software, 197
 - MEME Suite web server, 99
 - Metabolic modeling, 406
 - Metabolomics, 23, 124–125, 243, 495, 593
 - in agriculture, 153–154
 - application in, 13–14, 156–157
 - biochemical, nutritional, and toxicological features, 139
 - biotic and abiotic stresses assessment, 126–127, 126f
 - brown planthopper resistance in rice, 258
 - challenges of, 14
 - crops production, 129
 - crops quality, 129–130
 - databases, 142–144, 143f
 - drought stress responses in maize, 288–289
 - environmental and ecological metabolomics, 146–147
 - extraction methods in, 147–148, 147f
 - genetically modified (GM), 139
 - metabolic engineering in plants, 144–146
 - metabolic genome-wide association studies, 149–150
 - metabolic quantitative trait loci, 149
 - normal and stress conditions in plants, 155–156
 - nutritional/functional aspect, 139
 - omics tools, 154–155
 - phenotype and genotype, 139
 - plant biotechnology, 140–141, 140f
 - in plant metabolome, 150–151
 - postharvest crops science, 130–132, 131f
 - profiling, identification, and quantification, 144, 145f
 - for soils science and soil conservation, 127–129
 - technologies involvement, 141–142
 - workflow of analysis
 - data mining, annotation, and processing in, 152
 - sample preparation, 151–152, 151f
 - statistical tools and biomarker identification, 152–153
 - Metagenomics, 62
 - library preparation, 443
 - library screening
 - screening-based on function, 445–446
 - sequence-based screening, 443–445
 - rhizosphere analysis, 442f, 443–446
 - sample collection and isolation, 443
 - for sustainable agriculture, 446–449
 - workflow, 442f
 - Metal-induced stress, 156
 - Methyl-C sequencing, 561–562
 - 6-methyl-5-hepten-2-one (MHO), 131
 - MIAME, 41
 - Microarray hybridization, 9–10
 - Microarray technologies, 378–379, 508
 - Microbes, 449
 - Micronutrient deficiency, 14
 - MicroRNAs (miRNAs), 222
 - Microsatellites, 51
 - Minisequencing, 524
 - MiRNA expression-related QTLs (miR-eQTLs), 568–569
 - MISA perlscript, 423
 - Modeller, 99
 - ModRefiner, 200–201
 - Molecular markers, 456–458, 553
 - Morphological markers, 538
 - Mosaic, 424
 - “Movement Proteins” (MP), 426–427
 - Multidimensional protein identification technology, 111
 - Multilayer perceptron, 594
 - Multioptic data integration, 596–597
 - Multioptics, 234
 - MUSCLE program, 198
- N**
- Nandina domestica*, 422
 - Nanopore sequencing technology, 380
 - National Center for Biotechnology Information (NCBI), 32–33
 - Necrosis, 425–426
 - Nervana Neural Network Processor-T 1000, 653
 - Network connection, 575
 - Neurogenomics, 62
 - Neutron activation analysis, 15
 - New York Structural Genomics Research Consortium, 493, 496–497
 - Next-generation sequencing (NGS), 3, 47, 181–183, 329, 361–362, 373, 562
 - in agronomic advancements, 222
 - applications of, 42
 - bioinformatics tools, 43
 - citrus, 215
 - illumina sequencing, 41–42
 - olive, 215–216
 - peanut, 215
 - for plant virus. *See* Plant virus
 - rice, 212–214
 - tea, 216
 - NGS-based RNA sequencing (RNA-seq), 9
 - Noncoding RNA, 477–479
 - Nongel-based method
 - minisequencing, 524
 - TaqMan assay, 524, 524f
 - Novel RNAs, in plants
 - chimeric RNA, 185–186
 - circular RNA, 184–185
 - long noncoding RNA, 181–183
 - small RNA, 177–181
 - Nuclear magnetic resonance (NMR), 235
 - Nucleomics, 62
 - Nucleotide sequence, 268–269
 - Nutri-genomics, 495–496
- O**
- Oases, 379–380
 - OCT B-scan cross-sectional images, 620
 - Omics
 - computomics
 - applications, 16
 - challenges, 16
 - crop protection and improvement, 3
 - defined, 233
 - disciplines of, 4f
 - drought stress responses in maize
 - bioinformatics tools and databases, 289–290
 - genomics, 284–287
 - metabolomics, 288–289
 - phenomics, 289
 - proteomics, 288–289
 - transcriptomics, 287–288
 - genomics
 - applications, 9
 - applications of, 7–8
 - challenges, 9–10
 - challenges in agricultural field, 8
 - NGS technologies for, 6
 - transcriptomic techniques, 9
 - insect pesticide resistance, 3
 - integration of omics, 235
 - interdisciplinary techniques, 3
 - ionomics, 14–15
 - in machine learning, 31–32
 - metabolomics
 - application in, 13–14
 - challenges of, 14
 - next-generation sequencing (NGS)
 - technology, 3
 - phenomics
 - applications, 5–6
 - challenges, 6
 - plant herbicide tolerance, 3
 - proteomics
 - applications, 10–11
 - challenges of, 12
 - technologies, 11–12
 - single-omics, 236
 - One Health initiative, 23
 - OpenArray, 468
 - Operating system, 575–576
 - Organellar Genome Draw (OGDRAW), 423
 - Organelle genomes, 423
 - Organic farming, 449
- P**
- PacBio, 24, 445
 - Paenarthrobacter aurescens* TC1, 409
 - Pangenomics, 62, 212
 - PARE sequencing, 568
 - Partially fermented oolong tea, 361
 - Pathogenicity, 374
 - Pathway mapping, 30

- PCR-based markers, 513–514
 PCR based screening, 443
 PDBsum, 99
Pelargonium, 420
 Personal genomics, 62
 Pesticides, 625–626
 Pharmacogenomics, 597–598
 Phase-change memory (PCM), 654
 Phenolic compounds, 265–266
 Phenomics
 applications, 5–6
 challenges, 6
 drought stress responses in maize, 289
 Phosphorylation, 114
 Phylogenetic analysis, 526
 Physiological QTL (pQTL), 562
Phytophthora infestans fungus, 221
 Piwi-interacting RNAs (piRNAs), 222
 PLANN (Plastome Annotator), 423
 Plant betagalactosidases
 mechanism of action, 205–206
 MiBGA and TBG4, 201
 molecular evolution of, 198–200
 protein sequence features of, 192–197, 193*t*,
 197*f*, 198*f*
 substrate specificity of, 201–204, 204*t*, 205*t*
 three-dimensional structural characteristics
 of, 200–201
 Plant biology, 234
 Plant biotechnology, 3
 Plant breeding, 542
 Plant disease
 database creation, 621
 disease identification
 convolutional neural network, 622–623
 feature extraction and classification,
 621–622
 identification, 622
 severity estimation, 623
 visual symptoms of, 620
 Plant growth-promoting microbes, 449
 Plant metabolomics. *See* Metabolomics
 Plant-omics, 233–234
 Plant Phosphorylation database (P³DB),
 289–290
 Plant virus
 crop quality and quantity, 383
 ELISA, 383
 next-generation sequencing (NGS)
 application of, 387*t*, 391*t*
 Artichoke latent virus, 386
 by bioinformatics tools, 385–386
 biological indexing and molecular biology
 techniques, 386
 CRISPER-Cas9, 387
 development of, 383–385
 quarantine virus detection methods, 386
 plant–viroid interactions, 383
Platanus occidentalis, 422
 Polymerase chain reaction (PCR), 459–460,
 537
 Polyploid genome assemblies, 15–16
Pospiviroidae, 373
 Postharvest crops science, 130–132, 131*f*
 Potato (*Solanum tuberosum* L.)
 biotic/abiotic stresses, 341
 bulked segregant analysis, 343–344
 genetic vulnerability, 341
 GWAS, 343–344
 ionomics, 352
 metabolomics
 abiotic traits, 351
 biotic traits, 351
 quality traits, 351–352
 molecular markers, 342–343
 nucleic acid sequencing and information
 technology, 341
 omics resources and integration of
 technologies, 353–354, 353*t*
 phenomics, 352–353
 proteomics, 348–350, 349*t*
 abiotic stress, 350
 biotic stress, 348–350
 quality traits, 350
 quantitative trait loci mapping, 343–344
 transcriptomics, 344–348
 abiotic stress, 347
 biotic stress, 346–347
 miRNAs in, 348
 quality traits, 347–348
 whole-genome sequencing and resequencing,
 342
 Powdery mildew (PM) disease, 302
 Pre-mRNA transcripts, 368
 Processor, 575
 Programming language, 578–579
 Progressive Filtering of Overlapping Small
 RNAs (PFOR), 380
 Protein amino acid sequences, 594
 Protein coding genes, 421
 Protein profiling, 289
 Protein sequence analyses, 198–200
 Protein structure, 572–574
 Protein Structure Initiative (PSI), 491–492
 Proteomics, 23, 107–108, 242–243, 272,
 494–495, 593
 applications, 10–11
 brown planthopper resistance in rice,
 257–258
 challenges of, 12
 drought stress responses in maize, 288–289
 technologies, 11–12
Pseudomonas, 401, 407
 PSTVd sequence variants, 379
Puccinia triticina, 301
 Pyrosequencing, 384
- Q**
 Quality analysis (QA) filtering, 305
 Quality control, 27
 Quantitative PCR (qPCR), 28
 Quantitative trait loci
 advantages and disadvantages of, 550–551
 construction of genetic linkage maps, 548,
 550*t*
 detection, 550
 genetic distance and mapping functions, 550
 genomic variation to expression variation,
 565–566
 identification of polymorphism, 549
 linkage analysis of markers, 549
 mapping population, 549
 molecular markers, 548, 550*t*
 molecular marker system for genotyping,
 562–564
 RNA sequence, 564–565
 Quantitative trait loci (QTLs), 9, 253–254,
 268, 270, 295, 363, 456–458
 Quercetin, 265–266
- R**
 RAM, 575
 Random amplified polymorphic DNA (RAPD),
 459–460, 463, 513
 development of genetic markers, 542
 genetic mapping, 541
 plant breeding, 542
 population genetics, 542
 Random forest, 621
 RAST (Rapid Annotations using Subsystems
 Technology) algorithms, 409
 Recombinant inbred lines (RILs), 460,
 562–563
 Recurrent Neural Networks (RNN), 594
 RedCom, 407
 Reduced-representation sequencing (RRS)
 method, 330–331
 Reference genome sequence, 361–362, 514
 Reference sequence strategy, 514, 515*f*
 basic local sequence alignment tool
 (BLAST), 514–515
 library preparation, 516–517
 Mapping and Assembly with Qualities
 (MAQ), 514–515
 next-generation sequencing, 517
 quality control and alignment, 517
 sample preprocessing and DNA or RNA
 extraction, 516
 sequence search and alignment by hashing
 algorithm (SSAHA), 514–515
 Short Oligonucleotide Alignment Program
 (SOAP), 514–515
 SNP calling, 517
 ungapped alignment, 515
 Regulatory small RNAs
 detection of, 568
 discovery and annotation, 566–567, 567*f*
 natural variation in, 568
 with quantitative trait loci mapping,
 568–569
 Reinforcement learning, 647–648
 Remote sensing, 629–630
 Remote sensing data, 602
 Resistive random-access memory (ReRAM),
 654
 Restriction Association DNA sequencing
 (RADseq), 468
 Restriction fragment length polymorphism
 (RFLP), 463
 in back crossing, 541

- in comparative mapping, 540
 - in DNA fingerprinting, 539–540
 - genetic traits, 541
 - linkage mapping with, 540–541
 - markers, 328
 - in species identification, 540
 - Restriction Site Associated DNA (RAD), 563
 - RAD-sequencing (RAD-seq), 210
 - Reverse sample probing technique, 442
 - Rhizosphere, 439–440
 - Rice (*Oryza sativa* L.)
 - abiotic and biotic factors, 253
 - in brown planthopper resistance
 - bioinformatics, 259
 - genomics, 253–255, 255*t*, 256*t*
 - metabolomics, 258
 - proteomics, 257–258
 - transcriptomics, 256–257
 - functional genomics, 45
 - global rice production, 253
 - omics-aided analysis, 253
 - RNA databases, 44
 - RNA-directed DNA methylation (RdDM), 566
 - RNase H2 enzyme-based amplification, 466
 - RNA-Seq, 42, 163, 365, 367*f*, 378–379, 564–565
 - abiotic stress, 163
 - annotation and functional analysis, 170–171
 - crops, 163
 - data analysis, 169
 - data generation, 165–168
 - differential expression analysis, 170
 - gene expression data validation, 165
 - gene expression profiling, 163–165
 - identified manuscripts, 163–165, 164*f*
 - ncRNAs, 165
 - omics data, 165
 - physiological data, 165
 - plant tissues/organs, 163
 - qPCR assays, 165
 - quality of assembly, 169–170
 - raw data processing, 168–169
 - relative quantification method, 165
 - studied accession, 165
 - transcriptome assembly, 165
 - transcript quantification, 170
 - RNA silencing mechanism, 378–379
 - Robotic elevated phenotyping, 15–16
 - Robotics, 641–642
 - Roche 454 genome sequencer, 444–445
 - Root system architecture (RSA), 303
 - Rubber tree, 507–508
- S**
- Salicylic acid (SA) signaling, 254
 - Salinity stress, 155
 - Salt Overly Sensitive 2 (SOS2) gene, 574
 - SAMTOOLS/GATK tools, 25
 - Sanger sequencing, 162, 375, 383–384, 508, 563
 - SearchSmallRNA, 379–380
 - Semisupervised learning, 648–649
 - Sentinel-2 data, 608
 - Sequence based approaches, 286
 - Sequence characterize amplified region (SCAR), 328, 459–460, 464
 - Sequence tagged site (STS), 459–460
 - Sequence variants, 376
 - Sequencing-by-synthesis approach, 41
 - Sequencing technologies
 - big data storage and management, 32–33
 - epigenomics, 29
 - future directions, 33
 - genome assembly technology
 - genome-wide association, 26
 - postassembly algorithms, 25
 - reference-based and de novo assembly, 24–25
 - metabolomics, 30
 - multipronged approach, 23
 - noncoding RNA, 28
 - omics datasets, 30–31, 31*t*
 - proteomics, 29–30
 - RNA-seq data analysis
 - alignment, 27
 - differential expression, 27–28
 - quality control, 27
 - quantification, 27
 - validation of, 28
 - Sequential window acquisition of all theoretical mass spectra, 111
 - Serial analysis of gene expression (SAGE), 286, 505
 - abnormal genome, 38–39
 - advantages of, 39
 - 15-bp tag sequence, 38–39
 - databases, 39
 - drawbacks of, 39–40
 - expressed sequence tags (ESTs), 38–39
 - 5'-GGGAC-3 sequence, 38–39
 - normal gene structure, 38–39
 - procedure, 38*f*
 - USAGE, 39
 - Severity estimation, 623
 - Short Read Archive (SRA), 361–362
 - Simple sequence repeat (SSR), 361–362, 459–460, 464, 513, 562–563
 - distribution of, 545
 - isolation of, 545–546
 - microsatellite
 - comparative mapping, 546
 - functional diversity, 546
 - mapping of gene, 546
 - software for, 553–555
 - Single-cell RNA sequencing, 597
 - Single marker analysis, 563
 - Single-molecule real-time (SMRT) sequencing approach, 49
 - Single-nucleotide polymorphism (SNP), 6, 305, 459–460, 466, 513–514
 - application, 548
 - biallelic mapping, 514
 - database, 520–521, 521*t*
 - dbNSFP tool, 518–520
 - de novo sequence strategy, 514, 516*f*
 - high-quality raw read, 518, 519*t*
 - input datasets, 517–518
 - quality control, 518
 - Teraclu, 517–518
 - TGICL, 517–518
 - discovery of, 368
 - diversity array technology (DArT Seq), 548
 - gel-based SNP genotyping
 - cleaved amplified polymorphic sequence markers, 522, 522*f*, 523*f*
 - single-stranded conformation polymorphism, 522
 - gene ontology (GO), 518–520
 - genotyping, 237
 - GoGene, 518–520
 - Medical Subject Headings (MeSH)
 - vocabularies, 518–520
 - molecular genetic marker and the potential number, 514
 - nongel-based method
 - minisequencing, 524
 - TaqMan assay, 524, 524*f*
 - plants, application in
 - genetic diversity, 525
 - genetic mapping, 525–526
 - marker-assisted selection, 526–528
 - phylogenetic analysis, 526
 - reference sequence strategy, 514, 515*f*
 - basic local sequence alignment tool (BLAST), 514–515
 - library preparation, 516–517
 - Mapping and Assembly with Qualities (MAQ), 514–515
 - next-generation sequencing, 517
 - quality control and alignment, 517
 - sample preprocessing and DNA or RNA extraction, 516
 - sequence search and alignment by hashing algorithm (SSAHA), 514–515
 - Short Oligonucleotide Alignment Program (SOAP), 514–515
 - SNP calling, 517
 - ungapped alignment, 515
 - SNP detection, 547
 - Var2GO, 518–520
 - variant calling format (VCF), 518–520
 - in vitro techniques, 547–548
 - Single Strand Conformation Polymorphism, 376
 - Skip-Gram model, 594
 - S. lycopersicum*, 266–267
 - Small interfering (siRNA)-directed RNA, 419–420
 - Small interfering RNA (siRNAs), 222, 226
 - Small noncoding RNA (sncRNAs)
 - crops and livestock, 221
 - insect pests, 221
 - limitations, 226
 - types of, 222
 - Small RNA, 221–222, 379, 390–392, 391*t*
 - heterochromatic small-interfering RNA, 180
 - microRNA, 177–179, 179*t*, 222–225
 - natural antisense-small-interfering RNA, 180
 - PIWI-interacting RNAs, 225–226
 - phased small-interfering RNA, 180
 - small-interfering RNA, 179–180

- Small RNA (*Continued*)
 trans-acting small-interfering RNA, 180
 transfer RNA–derived small RNA, 181, 182*t*
- SMART (Simple Modular Architecture Research Tool), 99
- Smart agriculture and big data
 adaptation to climate change, 628
 automated irrigation system, 628
 digital soil and crop mapping, 627
 disease detection and pest management, 628
 fertilizers recommendation, 628
 weather prediction, 627
- Sodium dodecyl sulfate-polyacrylamide gel electrophoresis, 110
- Software, AI
 Amazon artificial intelligence services, 655
 Google artificial intelligence platform, 655
 Infosys Nia, 656
 Microsoft Azure, 655
 Rainbird, 656
 TensorFlow, 655
 Wipro HOLMES, 656
- Soil microbiota
 agricultural microbiomes, 439
 common culture-dependent and culture-independent methods, 440*f*, 441*f*
 culture-dependent methods, 439
 environmental equilibrium, 440
 metagenomics
 rhizosphere analysis, 442*f*, 443–446
 for sustainable agriculture, 446–449
 microbial diversity, 439
 rhizosphere, 440–441
 rhizosphere microbial community, 441–442
 soil microbial diversity, 441
 sustainable agriculture, 439–440
- Soil microorganism, 446
- Soils science and soil conservation, 127–129
- Solanum lycopersicum*, 265–266
- Sparse Autoencoder, 594
- 16S rRNA sequence strategy, 445
- Stable-isotope labeling, 288
 by amino acids in cell culture, 113
- Stem rust (SR) disease, 302
- Storage, 575
- Structural genomics (SG), 63
 atomic structure of proteins, 491–492
 gene and protein interactions, 493
 genome sequencing focuses, 491–492
 high-throughput (HT) strategies, 491–492
 single protein structures, 492–493
 three-dimensional structure, 493
- Substitution-mediated enhanced salt tolerance, 574
- Supervised learning, 646–647
- Support vector machine (SVM), 621
 regression, 16
- Sustainable agriculture, 123
 agricultural research resource allocation, 124
 and agro-production systems, 124
 bioinformatics, 455
 database resources for, 455–459
- breeding and genetics, emerging strategies for
 gene expression regulation by noncoding RNA, 477–479
 “omics” data, 479
 sustainable crop production, 479–480
 defined, 123
- DNA marker development and application
 amplified fragment length polymorphisms, 463–464
 automation and throughput genetic markers, 469
 breeding programs, 465
 cleaved amplified polymorphic sequences/derived cleaved amplified polymorphic sequences, 464
 high-throughput genotyping technologies, 467–469
 medium-throughput genotyping technologies, 465–466
 random amplified polymorphic DNA, 463
 restriction fragment length polymorphism, 463
 sequence characterized amplified region, 464
 simple sequence repeats, 464
 single-nucleotide polymorphism and insertion/deletion markers, 465
 single-nucleotide polymorphism genotyping, 469–474
- ecosystems and soils, 123–124
 environment quality, 123–124
 genome mapping
 comparative mapping, 461–462
 molecular marker systems and populations, 459–461
 physical mapping, and genome sequencing, 461
 practical applications of, 462
- genome-wide association studies
 design and analysis, 475–476
 to plant and animal breeding, 476
 single-nucleotide polymorphism markers, 474–475
- metabolomics and, 124–125
 biotic and abiotic stresses assessment, 126–127, 126*f*
 crops production, 129
 crops quality, 129–130
 postharvest crops science, 130–132, 131*f*
 for soils science and soil conservation, 127–129
 natural resources’ preservation, 123–124
- T**
- Tablet, 43
- Tandem mass tag (TMT), 112–113
- TaqMan assay, 524, 524*f*
- TaqMan SNP genotyping technology, 466
- Tea Plant Information Archive (TPIA), 361–362
- Tea transcriptome sequencing
 ABA biosynthesis-related genes, 368
 alternative splicing (AS), 368
 cDNA synthesis, 365
C. sinensis var *sinensis* (CSS), 366
 RNA-Seq methodology, 365, 367*f*
 seasonal variation, 368
 stress management of tea, 367*t*
 tissue-specific expressed genes, 366
- Temperature stress, 156
- Tensor Processing Unit (TPU), 652
- Theaflavins, 361
- The Food and Agriculture Organization of the United Nations, 449–450
- The Integrative Genomics Viewer, 43
- The International Rice Functional Genomics Steering Committee, 496
- Thermal infrared cameras, 4
- Third-generation sequencing, 380
- TILLING method, 269
- Tissue-specific expressed genes, 366
- Tn5 transposase, 571–572
- Tomato
 abiotic stress, 270–272
 breeding, 269
 disease resistance, 269–270
 genetic markers for selection, 271
 genome and genetic variation of, 267–269
 insect resistance, 270
 proteomics, 272
 transcriptomics, 272
- Tomato mosaic virus (ToMV), 426–427
- Tomato planta macho viroid* (TPMVD), 373
- TopHat, 43
- Trait-based approaches, 283
- Transcriptional control, 579
- Transcription factors (TFs)
 activation domain, 79
 bioinformatics tools
 BLAST tool, 97
 Clustal programs, 98
 data mining, 94–96
 expression analysis, 94*f*
 gene duplication and functional divergence studies, 100
 gene predictions, 100
 genome-wide analysis of transcription factors, 91*f*
 genomic data analysis, 91*f*
 Kalign algorithm, 98
 MAFFT, 98
 motif and domain prediction, 98–99
 MUSCLE, 98
 phylogenetic studies, 92*f*
 physicochemical properties analysis, 98
 in silico structure prediction of proteins, 99
 T-Coffee, 98
 tertiary structure predictions, 93*f*
 transcription factor gene families, 95*t*
 for biotic and abiotic tolerance, 84
 databases, 84
 plant transcription factors and multifarious applications
 AP2/ERF family, 80
 bHLH family, 81

- bZIP, 81
 DNA binding, 81–82
 MADS family, 82
 Myeloblastosis family, 82
 NAM/ATAF/CUC family, 82–83
 WRKY family, 83
 Zinc fingers, 83–84
 sequence-specific DNA-binding proteins, 79
 Transcriptome assembly, 65, 65*t*, 165
 Transcriptome sequencing (RNA-seq), 212
 Transcriptomics, 8–10, 23, 241, 272, 494, 593
 biochemical markers, 504
 brown planthopper resistance in rice, 256–257
 cytological markers, 504
 DNA sequence–based markers, 506
 drought stress responses in maize, 287–288
 expressed sequence tags, 505
 first-generation sequencing, 161
 GBS-t analysis, 506–507
 gene expression, 161
 gene expression cycles, 503
 gene regulatory events, 503
 genetic and transcriptome-based markers, 506
 genetic markers, 506–507
 genetic polymorphism, 507
 microarrays, 505
 microarray technology, 161–162
 molecular markers, 504
 nucleic acid sequencing cost, 163
 outbreeding and inbreeding crops, 507
 phenotypic markers, 503
 plants, markers in, 504
 reverse transcription (RT), 161
 RNA-Seq. *See* RNA-Seq
 RNA sequencing, 505–506
 RT protocol, 161
 Sanger-sequencing method, 162
 serial analysis of gene expression technology, 505
 single-nucleotide polymorphism (SNP) frequency, 506
 SuperSAGE method, 162
 transcriptomics studies, 162*f*
 Transcript quantification, 170
 Translational genomics, 240
 Transposable elements (TE), 566
 Trans-regulatory elements (TRE), 566
 T-tests, 563
- Two-dimensional-difference-in-gel electrophoresis, 110–111
 Two-dimensional gel electrophoresis (2DGE), 107–108, 110
- U**
- Unfermented green tea, 361
 Unsupervised learning, 647
 Unsupervised Machine Learning, 590
 UN Sustainable Development Goals, 455
 USEARCH, 379–380
- V**
- Var2GO, 518–520
 Variant calling format (VCF), 468, 518–520
 VCAKE, 379–380
 Vegetation indices, 633–635
 Velvet, 379–380
 Verdant, 423
 Viral replication complexes (VRCs), 426
 VirFind, 380
 Viroids
 disease incidence, 374
 high light intensity and high temperature, 374
 next-generation sequencing technology, 374–375
 bioinformatic intervention in, 379–380
 impact of, 375–376, 375*t*, 377*t*
 role of, 376–379
 (–) polarity RNA sequences, 373–374
 RNA virus, 373
 self-complementarity, 373
 Virtool, 380
 VirusDetect, 380
 Visualization, 423
- W**
- Waterlogging stress, 155
 Wavelet-based filtering, 609
 Weather prediction, 627
 Weighted gene coexpression network analysis (WGCNA), 30, 565
 Wheat (*Triticum aestivum* L.)
 comparative genomics, 295
 “conserved” DNA sequences, 295
 crop modeling studies, 295
 DNA-based molecular marker technology, 323
 gene discovery and marker development
 biofortification hotspots, 298–300
 for biotic stress resistance, 300–303
 colinearity-based gene cloning, 296–297
 for drought stress tolerance, 303–304
 functional comparative genomics, 298
 gene annotation and marker development, 297–298
 genomic hotspots for, 304
 genome-wide markers for gene mapping, 328
 genomic sequences, 305–313
 genomics for development of marker, 328–329
 genotyping-by-sequencing, 330–331
 geographic distribution and exposure, 323
 high gene plasticity, 323
 high-throughput genotyping approaches, 331–332
 microarray-based genotyping, 329–330
 molecular marker systems, 324–327, 325*t*, 326*t*
 quantitative trait loci (QTLs), 295
 trait-linked SNPs, 332–333
 Wheat660K SNP array, 330
 Whole genome sequencing, 209–210
 genome assembly process, 365*f*
 identification and characterization, 365
 predicted protein-coding genes, 363
 quantitative trait loci (QTLs), 363
 well-annotated genome and chromosome-scale tea genome, 363
 World Bank, 123
- X**
- X-beam crystallography, 491–493
 X-ray fluorescence (XRF) spectroscopy, 15
- Y**
- Yellows/Chlorosis, 425
- Z**
- Zeaxanthin, 266

Bioinformatics in Agriculture

Next-Generation Sequencing Era

Edited by **Pradeep Sharma, Dinesh Yadav and Rajarshi Kumar Gaur**

Harnesses genomics technologies for genetic engineering and pathogen characterization and diagnosis

Edited by

Pradeep Sharma

Indian Institute of Wheat and Barley Research, Karnal, India

Dinesh Yadav

Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, India

Rajarshi Kumar Gaur

Department of Biotechnology, D.D.U. Gorakhpur University, Gorakhpur, India

Bioinformatics in Agriculture: Next-Generation Sequencing Era is a comprehensive volume presenting an integrated research and development approach to the practical application of genomics to improve agricultural crops. Exploring both the theoretical and applied aspects of computational biology and focusing on the innovation processes, the book highlights the increased productivity of a translational approach. Presented in four sections and including insights from experts from around the world, the book includes the following: Section I: Bioinformatics and Next-Generation Sequencing Technologies; Section II: Omics Application; Section III: Data Mining and Markers Discovery; Section IV: Artificial Intelligence and Agribots.

This book explores deep sequencing, NGS, genomic, transcriptome analysis and multiplexing, highlighting practices for reducing time, cost, and effort for the analysis of genes as they are pooled, and sequenced. Readers will gain real-world information on computational biology, genomics, applied data mining, machine learning, and artificial intelligence.

This book serves as a complete package for advanced undergraduate students, researchers, and scientists with an interest in bioinformatics.

- Discusses integral aspects of molecular biology and pivotal tools for molecular breeding enables breeders to design cost-effective and efficient breeding strategies
- Provides examples of innovative genome-wide marker (SSR and SNP) discovery
- Explores both the theoretical and practical aspects of computational biology with a focus on innovation processes
- Covers recent trends of bioinformatics and different tools and techniques



ACADEMIC PRESS

An imprint of Elsevier

elsevier.com/books-and-journals

ISBN 978-0-323-89778-5



9 780323 897785