

Advances in Biochemical Engineering/Biotechnology 164
Series Editor: T. Scheper

Rajeev K. Varshney · Manish K. Pandey
Annapurna Chitikineni *Editors*

Plant Genetics and Molecular Biology

 Springer

164

**Advances in Biochemical
Engineering/Biotechnology**

Series editor

T. Scheper, Hannover, Germany

Editorial Board

S. Belkin, Jerusalem, Israel

T. Bley, Dresden, Germany

J. Bohlmann, Vancouver, Canada

M.B. Gu, Seoul, Korea (Republic of)

W.-S. Hu, Minneapolis, Minnesota, USA

B. Mattiasson, Lund, Sweden

J. Nielsen, Gothenburg, Sweden

H. Seitz, Potsdam, Germany

R. Ulber, Kaiserslautern, Germany

A.-P. Zeng, Hamburg, Germany

J.-J. Zhong, Shanghai, Minhang, China

W. Zhou, Shanghai, China

Aims and Scope

This book series reviews current trends in modern biotechnology and biochemical engineering. Its aim is to cover all aspects of these interdisciplinary disciplines, where knowledge, methods and expertise are required from chemistry, biochemistry, microbiology, molecular biology, chemical engineering and computer science.

Volumes are organized topically and provide a comprehensive discussion of developments in the field over the past 3–5 years. The series also discusses new discoveries and applications. Special volumes are dedicated to selected topics which focus on new biotechnological products and new processes for their synthesis and purification.

In general, volumes are edited by well-known guest editors. The series editor and publisher will, however, always be pleased to receive suggestions and supplementary information. Manuscripts are accepted in English.

In references, *Advances in Biochemical Engineering/Biotechnology* is abbreviated as *Adv. Biochem. Engin./Biotechnol.* and cited as a journal.

More information about this series at <http://www.springer.com/series/10>

Rajeev K. Varshney · Manish K. Pandey ·
Annapurna Chitikineni
Editors

Plant Genetics and Molecular Biology

With contributions by

V. Anil Kumar · J. Batley · P. Chaturvedi · A. Chitikineni ·
J. Cockram · R. R. Das · S. Datta · D. Edwards · A. Ghatak ·
J. Jankowicz-Cieslak · Y. Jia · K. Jiang · P. L. Kulwal ·
I. Mackay · N. Mantri · P. R. Marri · S. Mazicioglu ·
M. Muthamilarasan · N. Nejat · I. Ocsoy · G. Pandey ·
M. K. Pandey · S. K. Pandey · A. Parveen · M. Prasad ·
A. Ramalingam · C. S. Rao · A. Rathore · S. D. Rounsley ·
J. K. Roy · M. Saba Rahim · A. Scheben · H. Sharma ·
V. K. Singh · W. Tan · D. Tasdemir · V. Thakur · B. J. Till ·
R. K. Varshney · W. Weckwerth · C. B. Yadav · L. Ye

 Springer

Editors

Rajeev K. Varshney
International Crops Research Institute
for the Semi-Arid Tropics (ICRISAT)
Hyderabad, India

Manish K. Pandey
International Crops Research Institute
for the Semi-Arid Tropics (ICRISAT)
Hyderabad, India

Annapurna Chitikineni
International Crops Research Institute
for the Semi-Arid Tropics (ICRISAT)
Hyderabad, India

ISSN 0724-6145

ISSN 1616-8542 (electronic)

Advances in Biochemical Engineering/Biotechnology

ISBN 978-3-319-91312-4

ISBN 978-3-319-91313-1 (eBook)

DOI 10.1007/978-3-319-91313-1

Library of Congress Control Number: 2018948681

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The elimination of hunger and malnutrition from society is a key challenge of all agricultural stakeholders around the world. Feeding the global population has never been so challenging, especially in the context of diminishing land and water resources, an ever-increasing global population, and climate change. The only solution may be to develop climate-smart plant varieties that are produced with appropriate agricultural management practices. Today, agriculture is facing an acute shortage of advanced germplasms to replace inferior varieties in farmers' fields. A "game-changer" strategy for the development of improved germplasms and cultivation practices needs to be implemented quickly and precisely to tackle both current and future adverse environmental conditions.

Fast-evolving technologies can serve as a potential growth engine in agriculture because many of these technologies have revolutionized other industries in the recent past. The tremendous advancements in biotechnology methods, cost-effective sequencing technology, refinement of genomic tools, standardization of modern genomics-assisted breeding methods, and digitalization of the entire breeding process and value chain hold great promise for taking global agriculture to the next level through the development of improved climate-smart seeds. These technologies can dramatically increase our capacity for understanding the molecular basis of traits and utilizing the available resources for accelerated development of stable, high-yield, nutritious, efficient, and climate-smart crop varieties. These improved crop varieties and agricultural practices will help us to address global food security issues in an equitable and sustainable manner.

For these reasons, this book aims to explore and discuss future plans in the key areas of plant genetics and molecular biology. It contains 12 chapters written by 42 authors from Australia, Austria, India, Turkey, the United Kingdom, and the United States (see List of Contributors). The editors are grateful to all of the authors for contributing high-quality chapters with information from their areas of expertise. The editors also would like to thank the reviewers (see List of Reviewers) for their help in providing constructive suggestions and corrections, which helped the authors to improve the quality of the chapters. The editors are also

grateful to Dr. David Bergvinson (Director General, ICRISAT) and Dr. Peter Carberry (Deputy Director General—Research, ICRISAT) for their encouragement and support. The editors thank the series editors (T. Scheper, S. Belkin, T. Bley, J. Bohlmann, M.B. Gu, W.-S. Hu, B. Mattiasson, J. Nielsen, H. Seitz, R. Ulber, A.-P. Zeng, J.-J. Zhong and W. Zhou) of the Springer publication *Advances in Biochemical Engineering/Biotechnology* (<http://www.springer.com/series/10>) for giving us this opportunity to compile such a wealth of information on plant genetics and molecular biology for the research and academic community. The assistance received from Springer—in particular, Judith Hinterberg, Elizabeth Hawkins, Arun Manoj, and Alamelu Damodharan—has been a great help in completing this book. The cooperation and encouragement of the publisher are gratefully acknowledged.

We also appreciate the cooperation and moral support from our family members, especially when the precious time we should have spent with them was taken up by editorial work. R.K.V. acknowledges the help and support of his wife Monika, son Prakhar, and daughter Preksha, who allowed their time to be taken away to fulfill R. K.V.'s editorial responsibilities in addition to research and other administrative duties at ICRISAT. Similarly, M.K.P. is grateful to his wife Seema for her help and moral support during the evenings and weekends of editorial responsibilities in addition to research duties at ICRISAT, with special thanks to his brave daughter, the late Tanisha, who was alive for only a short period of time (3 months) after birth. A.C. thanks her husband Sudhakar and daughter Shruti for their cooperation and understanding during the fulfillment of her editorial commitments.

We hope that our efforts in compiling the information herein on the different aspects of plant genetics and molecular biology will help researchers to develop a better understanding of the subject and frame future research strategies. In addition, we hope that this book will also benefit students, academicians, and policymakers in updating their knowledge on recent advances in plant genetics and molecular biology research.

Hyderabad, India

Rajeev K. Varshney
Manish K. Pandey
Annapura Chitikineni

Contents

Plant Genetics and Molecular Biology: An Introduction	1
Rajeev K. Varshney, Manish K. Pandey, and Annapurna Chitikineni	
Advances in Sequencing and Resequencing in Crop Plants	11
Pradeep R. Marri, Liang Ye, Yi Jia, Ke Jiang, and Steven D. Rounsley	
Revolution in Genotyping Platforms for Crop Improvement	37
Armin Scheben, Jacqueline Batley, and David Edwards	
Trait Mapping Approaches Through Linkage Mapping in Plants	53
Pawan L. Kulwal	
Trait Mapping Approaches Through Association Analysis in Plants	83
M. Saba Rahim, Himanshu Sharma, Afsana Parveen, and Joy K. Roy	
Genetic Mapping Populations for Conducting High-Resolution Trait Mapping in Plants	109
James Cockram and Ian Mackay	
TILLING: The Next Generation	139
Bradley J. Till, Sneha Datta, and Joanna Jankowicz-Cieslak	
Advances in Transcriptomics of Plants	161
Naghmeh Nejat, Abirami Ramalingam, and Nitin Mantri	
Metabolomics in Plant Stress Physiology	187
Arindam Ghatak, Palak Chaturvedi, and Wolfram Weckwerth	
Epigenetics and Epigenomics of Plants	237
Chandra Bhan Yadav, Garima Pandey, Mehanathan Muthamilarasan, and Manoj Prasad	
Nanotechnology in Plants	263
Ismail Ocsoy, Didar Tasdemir, Sumeyye Mazicioglu, and Weihong Tan	

Current Status and Future Prospects of Next-Generation Data Management and Analytical Decision Support Tools for Enhancing Genetic Gains in Crops 277
Abhishek Rathore, Vikas K. Singh, Sarita K. Pandey, Chukka Srinivasa Rao, Vivek Thakur, Manish K. Pandey, V. Anil Kumar, and Roma Rani Das

Index 293

List of Contributors

V. AnilKumar International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Jacqueline Bately University of Western Australia, Crawley, WA, Australia

Palak Chaturvedi University of Vienna, Vienna, Austria

Annapurna Chitikineni International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

James Cockram National Institute of Agricultural Botany (NIAB), Cambridge, UK

Roma Rani Das International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Sneha Datta International Atomic Energy Agency (IAEA), Vienna, Austria

David Edwards University of Western Australia, Crawley, WA, Australia

Arindam Ghatak University of Vienna, Vienna, Austria

Joanna Jankowicz-Cieslak International Atomic Energy Agency (IAEA), Vienna, Austria

Yi Jia Dow Agrosciences, Indianapolis, IN, USA

Ke Jiang Dow Agrosciences, Indianapolis, IN, USA

Pawan L. Kulwal Mahatma Phule Agricultural University, Rahuri, India

Ian Mackay National Institute of Agricultural Botany (NIAB), Cambridge, UK

Pradeep R. Marri Dow Agrosciences, Indianapolis, IN, USA

Sumeyye Mazicioglu Erciyes University, Kayseri, Turkey

Mehanathan Muthamilarasan National Institute of Plant Genome Research (NIPGR), New Delhi, India

Naghmeb Nejat RMIT University, Melbourne, VIC, Australia

Ismail Ocsoy Erciyes University, Kayseri, Turkey

Garima Pandey National Institute of Plant Genome Research (NIPGR), New Delhi, India

Manish K. Pandey International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Sarita K. Pandey International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Afsana Parveen National Agri-Food Biotechnology Institute (NABI), Mohali, India

Manoj Prasad National Institute of Plant Genome Research (NIPGR), New Delhi, India

M. Saba Rahim National Agri-Food Biotechnology Institute (NABI), Mohali, India

Chukka Srinivasa Rao International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Abhishek Rathore International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Steve D. Rounsley Genus plc, De Forest, WI, USA

Joy K. Roy National Agri-Food Biotechnology Institute (NABI), Mohali, India

Armin Scheben University of Western Australia, Crawley, WA, Australia

Himanshu Sharma National Agri-Food Biotechnology Institute (NABI), Mohali, India

Vikas K. Singh International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Weihong Tan University of Florida, Gainesville, FL, USA

Didar Tasdemir Erciyes University, Kayseri, Turkey

Vivek Thakur International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Bradley J. Till International Atomic Energy Agency, Vienna, Austria

Rajeev K. Varshney International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Wolfram Weckwerth University of Vienna, Vienna, Austria

Chandra Bhan Yadav National Institute of Plant Genome Research (NIPGR),
New Delhi, India

Liang Ye Dow Agrosciences, Indianapolis, IN, USA

List of Reviewers

Harsha Gowda Institute of Bioinformatics (IoB), Bangalore, India

Himabindu Kudapa International Crops research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Chikelu Mba Food and Agriculture Organization (FAO), Rome, Italy

Reyazul Rouf Mir Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir (SKUAST-K), Sopore, India

Manish K. Pandey International Crops research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Lekha Pazhamala International Crops research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Samir Sawant CSIR-National Botanical Research Institute (NBRI), Lucknow, India

Vikas Singh International Rice Research Institute (IRRI) -South Asia Hub, Hyderabad, India

Mahendar Thudi International Crops research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

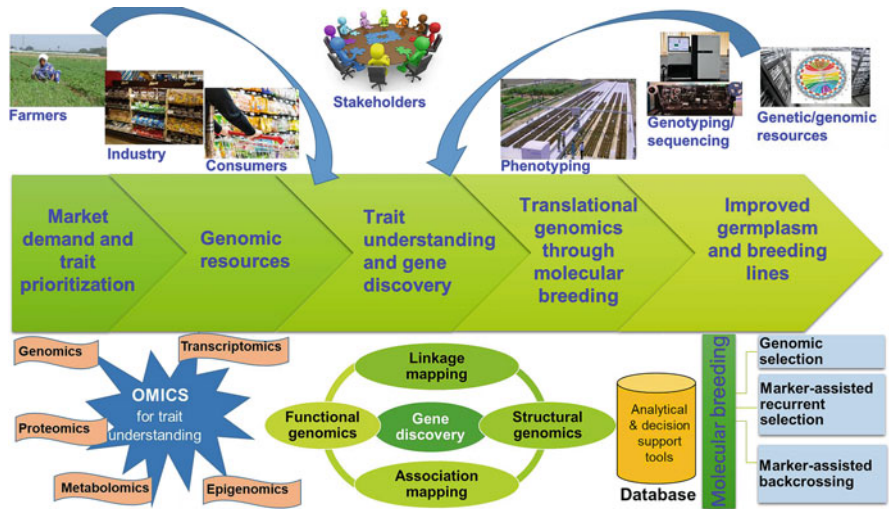
Plant Genetics and Molecular Biology: An Introduction



Rajeev K. Varshney, Manish K. Pandey, and Annapurna Chitikineni

Abstract The rapidly evolving technologies can serve as a potential growth engine in agriculture as many of these technologies have revolutionized several industries in the recent past. The tremendous advancements in biotechnology methods, cost-effective sequencing technology, refinement of genomic tools, and standardization of modern genomics-assisted breeding methods hold great promise in taking the global agriculture to the next level through development of improved climate-smart seeds. These technologies can dramatically increase our capacity to understand the molecular basis of traits and utilize the available resources for accelerated development of stable high-yielding, nutritious, input-use efficient, and climate-smart crop varieties. This book aimed to document the monumental advances witnessed during the last decade in multiple fields of plant biotechnology such as genetics, structural and functional genomics, trait and gene discovery, transcriptomics, proteomics, metabolomics, epigenomics, nanotechnology, and analytical tools. This book will serve to update the scientific community, academicians, and other stakeholders in global agriculture on the rapid progress in various areas of agricultural biotechnology. This chapter provides a summary of the book, “Plant Genetics and Molecular Biology.”

Graphical Abstract



Keywords Decision support tools, Epigenomics, Genomics, Metabolomics, Nanotechnology, Plant biotechnology, Proteomics, Transcriptomics

Contents

1 Introduction 2

2 High-Throughput Genotyping Platforms 4

3 Trait Dissection and Gene Discovery 5

4 Beyond Genomics 6

5 Data Management and Analytical Decision Supporting Tools 8

6 Summary 8

References 9

1 Introduction

Making society hunger-free and malnutrition-free is the main goal for the stakeholders in world agriculture. Feeding the global population has never been so challenging, especially in the context of diminishing land and water resources together with an ever-increasing global population and climate changes. One of the possible solutions is to develop climate-smart varieties of plants complimented with appropriate agricultural management practices. Today world agriculture is facing an acute shortage in developing improved germplasm to replace the old varieties existing in farmers’ fields. The global agriculture needs a “game-changer” strategy to be implemented with high priority in order to develop improved

germplasm and cultivation practices rapidly and with high precision to tackle the current and future adverse environmental conditions. Improved crop varieties together with improved agricultural practices will be able to address the global food security issue in an equitable and sustainable manner.

A recent survey on hunger and malnutrition has identified 52 of 119 countries as having a serious, alarming, or extremely alarming situation. Even today, 13% of the global population is undernourished and 27.8% of children under 5 years of age are stunted (<http://www.globalhungerindex.org/pdf/en/2017.pdf>). Despite the availability of sufficient food production, these problems still exist as a large number of people do not have access to nutritious food. The quality and nutrition of food products define the physical and mental health of the global population, not the quantity. In this context, agricultural research on developing nutrition-rich crops should be given equal importance to the major objective of increasing productivity. The genetic gains achieved over the decades in several crop species have been able to feed starving populations and have saved the lives of millions of people. Food and nutritional security in the coming years can only be made possible by achieving rapid and higher genetic gains in food crops with enhanced quality, nutrition, and adaptation to adverse climatic conditions. This goal can be achieved by integrating available biotechnological interventions with ongoing efforts. Not only agriculture but also biotechnology has been a great support in boosting several sectors such as the pharmaceutical, medical, and food processing sectors. In fact, the biotechnology interventions have already produced game-changing contributions in agriculture and the future contributions from biotechnology for society depend on strong policy, commitment, and the investment made in biotechnology research in coming years.

The rapid advances in biotechnological processes, approaches, and technologies have revolutionized agricultural research by developing a better understanding of plant genomes, gene discovery, genomic variations, and manipulation of desired traits in plant species. Additionally, these approaches also help researchers in developing a better understanding beyond genomes such as plant-pathogen and plant-environment interactions. The advanced technology support has helped to track the entire journey from genomes to phenotype using different “omics” approaches such as genomics (DNA/genome/genes), epigenomics (epigenetic modifications on the genetic material), transcriptomics (transcripts/RNA), proteomics (proteins), metabolomics (metabolites), interactomics (protein interactions), and phenomics (phenotype) (Fig. 1). The other important intervention is nanobiotechnology (a combination of nanotechnology and biology), which provides very sophisticated technical approach/devices for tracking, understanding, and solving biological problems. This book aimed to document current updates and advances in these frontier areas of biotechnology research. This chapter provides an overview of the different chapters included in the book.

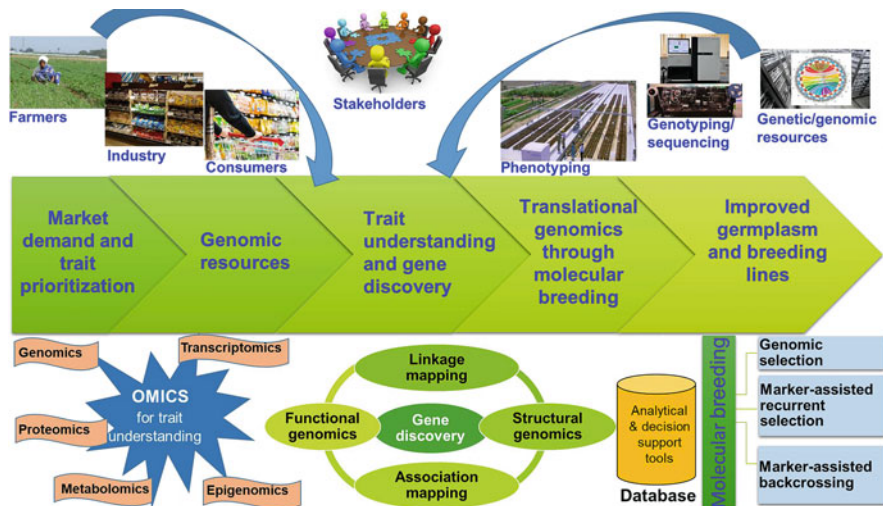


Fig. 1 Plant genetics and molecular biology for trait dissection and crop improvement

2 High-Throughput Genotyping Platforms

The tremendous advances in sequencing technologies have made it possible to sequence complete genomes of plant species for better understanding of the genome architecture evolution including whole genome duplications, dynamics of transposable elements, and several other components of the genome that define and control genome function leading to a particular phenotype [1]. Chapter 2 on “*Advances in Sequencing and Resequencing in Crop Plants*,” authored by SD Rounsley and other colleagues from Dow Agrosciences, USA and Genus plc, UK, provides updates on advancements in different sequencing technologies over the last two decades and their impact on plant genomics research. Cost-effective sequencing technologies have facilitated sequencing of a large number of plant genomes, which have impacted greatly on developing better understanding of plant genomes and their evolution [1, 2]. These advances have further helped in faster gene discovery, characterization, and deployment in plant improvement [3]. In addition to this, this chapter discusses the current challenges and future opportunities in further exploiting genomics information for plant improvement.

The reference genome of any plant species provides the foundation for genomics research, but mere sequencing of only one genome is not enough for harnessing the wealth of genetic diversity available within and across plant species. Therefore, sooner or later genome sequences will eventually be available for all the germplasm and exist in different genebanks for capturing the sequence variations followed by their manipulations using appropriate genetic improvement approaches such as

molecular breeding, genetic engineering (transgenics), genome editing, and any other such technology developed in future. Sequence variations in different genomes of the same species have been exploited as genetic markers for conducting different genetics and breeding studies.

Chapter 3 on “*Revolution in Genotyping Platforms for Crop Improvement*,” authored by David Edwards and his colleagues from the University of Western Australia (UWA), Australia, describes how different types of genetic variations can be used in genetics research and breeding applications through different genotyping platforms. Similar to sequencing, genotyping platforms have also gone through a rapid evolution and played an important role in advancing crop genetics and breeding. These genotyping platforms have been deployed in a range of genetic and breeding applications in most of the plant species. This chapter not only provides details on the evolution of different genotyping platforms over the decades, but also compares different genotyping platforms and predicts the future of genotyping in plants. This chapter clearly advocates the sequencing of entire genetic and breeding populations in future crop improvement programs for more precise and efficient plant selection in field.

3 Trait Dissection and Gene Discovery

The availability of genetic diversity is crucial for further improving the existing cultivars, which can sustain higher productivity under ever-challenging environments by acting as a buffer for adaptation and fighting climate change [4]. The development of improved cultivars using the diverse germplasm has helped farmers to replace these cultivars with older released or local varieties. The faster replacement of improved cultivars in the farmer’s field will help in achieving higher productivity under changing environments. Genomics-assisted breeding (GAB) holds great promise for accelerated development of improved cultivars; however, information on genes and diagnostic markers is required for deployment in any plant species. There are three major approaches of trait mapping, namely linkage mapping, linkage disequilibrium mapping/genome-wide association study (GWAS), and joint-linkage association mapping (JLAM).

Linkage mapping uses bi-parental genetic populations for traits with high variability between the parental genotypes. Chapter 4 on “*Trait Mapping Approaches through Linkage Mapping in Plants*,” authored by Pawan Kulwal from Mahatma Phule Agricultural University (MPAU), India, discusses different types of bi-parental populations and software for genetic mapping and quantitative trait locus (QTL) analysis in several plant species. Detailed information on key factors affecting the precision and accuracy of QTL discovery is presented. This mapping approach has been the most successful as diagnostic markers could be developed and deployed in breeding in several crop plants and many of these improved cultivars are grown in farmers’ fields.

In contrast to linkage mapping, the second trait mapping approach, genome-wide association study/linkage disequilibrium mapping, uses the diverse set of germplasm (natural population) and, therefore, no time is spent on development of genetic populations. The other advantage is that the association mapping panel can be used for mapping for several traits, while linkage mapping is possible for a couple of traits in a single bi-parental population. Furthermore, in many of the plant species, the development of bi-parental populations is not feasible or possible.

Chapter 5 on “*Trait Mapping Approaches through Association Analysis in Plants*,” authored by Joy Roy and his colleagues from the National Agri-Food Biotechnology Institute (NABI), India, provides greater insights different technical and applied aspects of GWAS analysis, advantages, and disadvantages of different software, and key factors affecting the precision and accuracy of results. This mapping approach has been deployed in many plant species.

The above two trait-mapping approaches have certain limitations and, therefore, the joint linkage association mapping approach came into existence; this approach can harness the advantages of both trait-mapping approaches. In this context, the shift now has moved from bi-parental to multi-parental populations, which allow high recombination leading to greater resolution for trait dissection. James Cockram and Ian Mackay from the National Institute of Agricultural Botany (NIAB), UK, in chapter 6 on “*Genetic Mapping Populations for Conducting High Resolution Trait Mapping in Plants*” summarize in-depth information on development and deployment of multi-parent populations such as multi-parent advanced generation intercross (MAGIC) and nested association mapping (NAM). This chapter also provides examples that showed better results in trait mapping in larger population size than in smaller ones.

All three above trait-mapping methods for trait mapping are forward genetics approaches, while Targeting Induced Local Lesions IN Genomes (TILLING) is a reverse genetics approach [5]. The TILLING approach involves creation of genetic variation through mutagenesis and then identification of genomic variation causing a change in phenotype. Chapter 7 on “*TILLING: The Next Generation*,” authored by Bradley Till and his colleagues from International Atomic Energy Agency (IAEA), Austria, describes the entire process of developing and deploying TILLING population for trait dissection and gene discovery. The chapter also discusses how integration of NGS technologies with TILLING have greatly accelerated the process of gene discovery. These populations also serve as a very good source for breeding and functional genomics studies.

4 Beyond Genomics

Genome sequencing greatly helped in understanding of genome organization and gene(s) structure that determines the basic features of each species. Nevertheless, just having genes in its genome does not provide certainty about the expected phenotype, which depends hugely upon other aspects of gene regulation. The

journey of a gene to a particular phenotype is very complicated, depending on as and when the DNA passes through different levels of regulation following the central dogma. It is, therefore, very essential to see beyond genomics for better clarity on gene function, networks, and interactions. In this context, the other “omics” approaches such as transcriptomics, proteomics, metabolomics, and interactomics play important roles in gene function and phenotype development. The phenotype is also affected by non-genomic elements, which bring epigenetic modifications to the genetic material, called as epigenomics. The epigenomic compounds modify the function of DNA without changing the sequence, thereby deviating from following the instruction of the genome. The interesting part is that these epigenetic features are being passed down over generations.

Transcriptomics plays an important role in gene discovery and functional characterization of the gene and its network. Chapter 8, authored by Nitin Mantri and his colleagues from RMIT University, Australia, on “*Advances in Transcriptomics of Plants*” discusses in detail discovery of transcriptional regulatory elements and deciphering mechanisms underlying transcriptional regulation. This chapter also covers related important aspects of gene regulation such as RNA splicing, microRNAs, small interfering RNAs (siRNAs), and long non-coding RNAs in plant development and response to biotic and abiotic stresses.

Metabolomics is very complex to understand due to development and interaction of the large number of metabolites produced during attaining metabolic homeostasis and biological balance in response to multiple cellular and extra-cellular factors. Wolfram Weckwerth and his colleagues from the University of Vienna, Austria, in chapter 9 on “*Metabolomics in Plant Stress Physiology*,” describe the importance of the study of metabolomics for functional genomics and system biology research leading to functional annotation of genes and better understanding of cellular responses for different biotic and abiotic stresses in plants. This chapter also provides details on different modern techniques that play a key role in developing more precise and high throughput data for comprehensive analysis. In addition to the above, this chapter also describes the complete processes involved in metabolomics study and lists the limitations faced by this scientific stream.

The epigenetic marks modifying the function of the gene can pass on over generations, making epigenomics an important component in better understanding the phenotype development. In other words, mere genome sequence is not responsible for phenotype development, and the epigenetic modifications play a key role by altering the chromatin structure and forcing deviation from the instructions contained in the genome. Detailed information on the types of epigenetic changes and their impact on phenotype development in plants is provided in chapter 10, entitled “*Epigenetics and Epigenomics of Plants*,” authored by Manoj Prasad and his colleagues from the National Institute of Plant Genome Research (NIPGR), India. This chapter also discusses the key role of NGS technologies and improved analytical software in better understanding the role of epigenomics in plant development and defense. Further information is also provided on different types of studies conducted in plants for identifying epigenetic factors and their potential role in plant improvement.

Nanotechnology has emerged recently as a very useful approach for plants and has already demonstrated its potential in the development of several nanomaterials in the pharmaceutical industry and in improving human health. Plants are the best source for developing such nanomaterials due to their large-scale availability and ease of production. Chapter 11 on “*Nanotechnology in Plants*,” authored by Ismail Ocoy and Weihong Tan and their colleagues from Erciyes University, Turkey and University of Florida, USA, explains the importance of nanotechnology in plants by citing several successful examples in medicine and industrial applications. The chapter mentions several advantages of plant extract over other biomolecules such as protein, enzyme, peptide, and DNA followed by their use in food, medicine, nanomaterial synthesis, and biosensing. This chapter also provides information on different extract preparation techniques, their use in the synthesis of nanoparticles, and demonstration of their antimicrobial properties against pathogenic and plant-based bacteria.

5 Data Management and Analytical Decision Supporting Tools

Large-scale data are generated at each step of the plant experiment related to understanding of the genome, gene discovery, functional characterization of gene, marker discovery, and deployment of diagnostic markers in the breeding program in addition to phenotyping data. All these data sets require efficient and effective database management systems, and analytical and decision support tools for storing and retrieving useful information that impacts the genetic improvement efforts. Chapter 12 on “*Current Status and Future Prospects of Next-generation Data Management and Analytical Decision Support Tools for Enhancing Genetic Gains in Crops*,” authored by Abhishek Rathore and his colleagues from ICRISAT, India, provides details on data management and analysis and decision support tools (DMAST) for plant improvement. The chapter also provides examples of how DMAST has simplified and empowered researchers in data storage, data retrieval, data analytics, data visualization, and sharing.

6 Summary

Ensuring food and nutritional security for an ever-increasing global population under the changing global climate is a top priority for policy makers across the globe. The existing conventional research efforts and traditional technologies will not be able to provide adequately nutritious food for the global population, necessitating the incorporation of modern science into the current genetic improvement programs. Biotechnology has great potential in bridging the supply-demand gap in

food through developing improved agricultural technologies. All the scientific streams are witnessing a rapid pace of development due to integration of new technologies such as robotics, automation, etc. These advancements have improved our understanding of genome architecture and its complexity: gene structure, function, and interactions, and improved methodologies for modification of the genome/gene to achieve a desired phenotype. The plant-pathogen and plant-environment interactions complicate the expression of scripts in the plant genome. This book covers these important research areas pertaining to plant biotechnology, which are key for achieving higher genetic gains. This wealth of information will be a great value for students, researchers, academicians, and policymakers.

References

1. Wendel JF, Jackson SA, Meyers BC, Wing RA (2016) Evolution of plant genome architecture. *Genome Biol* 17:37
2. Michael TP, Jackson S (2013) The first 50 plant genome. *Plant Genome* 6(2). <https://doi.org/10.3835/plantgenome2013.03.0001> in
3. Varshney RK, Nayak SN, Jackson S, May G (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 27(9):522–530
4. Buchanan-Wollaston V, Wilson Z, Tardieu F, Beynon J, Denby K (2017) Harnessing diversity from ecosystem to crop to genes. *Food Energy Secur* 6(1):19–25
5. Henikoff S, Till BJ, Comai L (2004) TILLING: traditional mutagenesis meets functional genomics. *Plant Physiol* 135(2):630–636

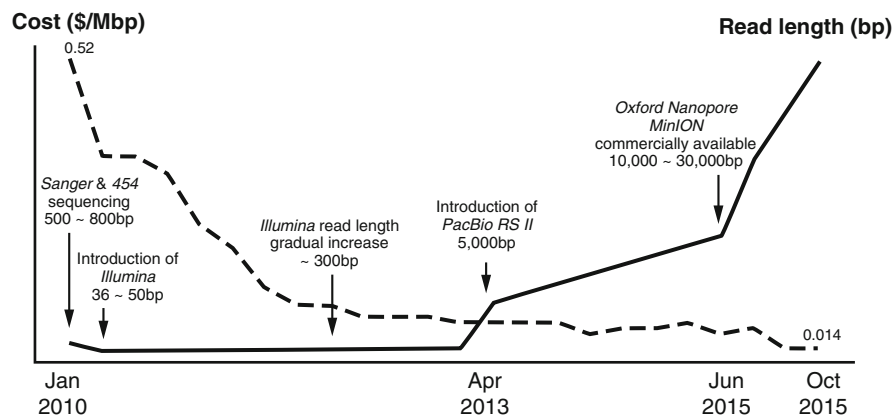
Advances in Sequencing and Resequencing in Crop Plants



Pradeep R. Marri, Liang Ye, Yi Jia, Ke Jiang, and Steven D. Rounsley

Abstract DNA sequencing technologies have changed the face of biological research over the last 20 years. From reference genomes to population level resequencing studies, these technologies have made significant contributions to our understanding of plant biology and evolution. As the technologies have increased in power, the breadth and complexity of the questions that can be asked has increased. Along with this, the challenges of managing unprecedented quantities of sequence data are mounting. This chapter describes a few aspects of the journey so far and looks forward to what may lie ahead.

Graphical Abstract



P. R. Marri, L. Ye, Y. Jia, and K. Jiang
Dow AgroSciences, Indianapolis, IN, USA

S. D. Rounsley (✉)
Genus plc, De Forest, WI, USA
e-mail: steve.rounsley@gmail.com

Keywords Assembly, Crops, NGS, Sequencing

Contents

1	Introduction	13
2	Current Technologies, Standards, and Strategies	14
2.1	Sequencing Technologies	14
2.2	Assembly Technologies	15
2.3	Reference Genome Project Strategies	17
2.4	Resequencing Strategies	20
2.5	Data Management and Visualization	20
3	Trends, Advanced Technologies, and Strategies	27
3.1	Sequencing Technologies	27
3.2	Assembly Strategies/Technologies	29
3.3	Genome Project Strategies	29
3.4	Resequencing Strategies	30
3.5	Data Management, Visualization, and Storage	30
3.6	Beyond Individual Variants: Alleles, Haplotypes, LD Blocks, and Pan-Genomes ...	30
4	Conclusion and Outlook	32
	References	32

Abbreviations

ABYSS	Assembly by Short Sequences
AGI	Arabidopsis Genome Initiative
API	Application Programming Interface
BAC	Bacterial Artificial Chromosome
CCD	Charge Coupled Device
CIGAR	Concise Idiosyncratic Gapped Alignment Report
CNV	Copy Number Variation
CRT	Cyclic Reversible Termination
DBG	de Bruijn Graph
ddNTPs	Dideoxynucleotides
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotides
GB	Giga-basepairs
GMOD	Generic Model Organism Database
GWAS	Genome Wide Association Mapping
HapMap	Haplotype Map
IGV	Integrative Genomics Viewer
InDel	Insertion-Deletion
Kb	Kilo-basepairs
LD	Linkage Disequilibrium
MAGIC	Multiparent Advanced Generation InterCross
Mb	Mega-basepairs

MTP	Minimum Tiling Path
NAM	Nested Association Mapping
NGS	Next-Generation Sequencing
OLC	Overlap Layout Consensus
ONT	Oxford Nanopore
PacBio	Pacific Biosciences
PAV	Presence-Absence Variation
PCAP	Parallel Contig Assembly Program
PCR	Polymerase Chain Reaction
PHRAP	Phil's Revised Assembly Program
PHRED	Phil's Read Editor
SBL	Sequencing by Ligation
SBS	Sequencing by Synthesis
SMRT	Single Molecule Real Time
SNA	Single Nucleotide Addition
SNP	Single Nucleotide Polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
Tb	Tera-basepairs
TIGR	The Institute for Genomic Research
UCSC	University of California at Santa Cruz
VCF	Variant Call Format
VEP	Variant Effect Predictor
WGS	Whole Genome Shotgun
ZMV	Zero Mode Waveguide

1 Introduction

When History of Science books are written in the future, there seems to be a more-than-reasonable chance that DNA sequencing and the birth of genomics will feature prominently. It is hard to think of a technology that has had a more dramatic effect on the study of biology than DNA sequencing. For those active in research today, with all the data and technology available, it is also hard to remember how little we knew about genomes before the mid 1990s. And despite the huge gulf in technology and knowledge between then and now, the field may still be in its infancy – in the first stages of a journey with a double helix as its guide. This chapter describes a few aspects of the journey so far and looks forward to what may lie ahead.

2 Current Technologies, Standards, and Strategies

2.1 Sequencing Technologies

2.1.1 Sanger Sequencing

In 1977, Frederick Sanger published a DNA sequencing technique that became the base technology for the field of genomics [1]. Sanger sequencing relies on the chain terminating properties of dideoxynucleotide triphosphates (ddNTPs), which were added to a mix of the four standard deoxynucleotides (dNTPs). When a complementary strand of sequence is synthesized using these reagents (the sequencing reaction), the result is a mixture of DNA fragments each terminated at different lengths. These fragments must then be separated by size (via electrophoresis), detected, and then recorded. Initially, slab polyacrylamide gels, radioactivity, and typing in sequence were integral to the standard (very manual) technique. Automated DNA sequencers were later developed, which automated the detection and capture of the resulting DNA sequence. Improvements such as fluorescently-labeled terminating nucleotides and capillary electrophoresis were incorporated into the ABI line of DNA sequencers. Hundreds of these instruments were sold to large genome centers working on genome projects in the 1990s and early 2000s – including bacteria, yeast, Arabidopsis, mouse, and human genomes [2].

2.1.2 Next-Generation Sequencing (NGS) Technologies

Over the last decade, sequencing technologies have evolved rapidly and led to a significant increase in throughput and reduction in cost, thereby enabling large-scale sequencing of genomes. They have done so by removing a limitation of Sanger sequencing of having to separate DNA fragments by size. In Sanger sequencing, the sequencing reaction occurs outside of the instrument, and the instrument simply separates and detects fragments. For most NGS technologies, the sequencing reaction is occurring on the instrument, and each base addition onto a growing DNA molecule is detected and recorded. The first generation of NGS technologies have relied largely on two approaches for sequencing, sequencing by ligation (SBL) and sequencing by synthesis (SBS) [3]. Both approaches rely on spatially constrained, clonal amplification of DNA and facilitate massive parallelization of sequencing reactions, each with its own clonal DNA template, resulting in the sequencing of millions of sequences in parallel.

SBL involves hybridization and ligation of fluorophore-labelled probes and anchor sequences to a DNA strand and capturing the emission spectrum to identify the DNA base, whereas SBS relies on strand extension using a DNA polymerase and uses changes in color or changes in ionic concentration to identify the incorporated nucleotide [3]. SBL is used in platforms such as SOLiD and Complete Genomics, whereas 454, Ion Torrent and Illumina use the SBS approach.

The SBS technologies can be classified into two approaches: the first, single nucleotide addition (SNA), used in 454 and Ion Torrent sequencers. This approach adds four nucleotides iteratively and scans for a signal after each to record an incorporated nucleotide. In the case of 454, which sold the first NGS instrument (the GS20), template-bound beads are distributed into a PicoTiterPlate and emulsion PCR is performed to clonally amplify a single DNA fragment within a water-in-oil microreactor. The addition of dNTPs triggers an enzymatic reaction that results in a fluorescent signal that is captured by a charge-coupled device (CCD) camera and is indicative of incorporated nucleotide [4]. The SNA method as implemented in Ion Torrent relies on ion sensing rather than fluorescence and detects the H^+ ions that are released after the incorporation of each dNTP and the resulting shift in pH is used to determine the incorporated nucleotide. Both 454 and Ion Torrent methods have limitations in accurately measuring the homopolymer lengths, because all nucleotides in a homopolymer are incorporated at the same time, and the magnitude of the signal must be used to estimate the homopolymer's length.

The other SBS approach is found in the NGS instruments that have come to dominate the market – those manufactured by Illumina. This technology, which was developed by Solexa before they were acquired by Illumina, uses terminating nucleotides similar to Sanger, except the termination is reversible. Cyclic reversible termination (CRT) uses a mixture of four reversible terminators each with a distinct fluorescence. Each template is extended by a single base only using the appropriate terminator and the resulting labeled templates are imaged recording which nucleotide was added to each template. The terminators are then cleaved off, and the cycle continues with the addition and imaging of the next nucleotide. An additional key to Illumina's success is the massive number of templates the technology can sequence in parallel – approaching three billion on a single flow cell in the HiSeq-X instrument. They achieve this through the immobilization of a DNA library onto a glass flow cell coated with adapter oligos. Clonal clusters of each DNA fragment are synthesized using bridge amplification on the flow cell resulting in a very large number of sequence-ready templates. Illumina currently has the largest market share for sequencing instruments and offers a wide variety of sequencing systems, read lengths, and throughput to cater to a wider range of applications (Table 1).

2.2 Assembly Technologies

The developments in automated, higher throughput sequencing technologies have been matched by concomitant development of algorithms and tools to use the resulting data in various applications. For projects where the goal is the generation of a reference genome, assembly algorithms have been a key area of development. The selection of an appropriate algorithm depends on the sequencing strategy being used (see next section), but here we will describe the main classes available.

Assembly algorithms can be broadly divided into two classes: overlap-layout-consensus (OLC) and De-Bruijn-graph (DBG) [5]. The OLC approach identifies

Table 1 Illumina sequencing systems

Metrics	MiniSeq	MiSeq v3	NextSeq	HiSeq2500 v4	HiSeq3000/4000	HiSeq X
Maximum output	7.5 Gb	15 Gb	120 Gb	1,000 Gb	1,500 Gb	1,800 Gb
Cluster number (millions)	25	25	400	4,000	5,000	6,000
Read length	1 × 75 bp	2 × 75 bp	1 × 75 bp	1 × 36 bp	1 × 50 bp	2 × 150 bp
	2 × 75 bp	2 × 300 bp	2 × 75 bp	2 × 50 bp	2 × 75 bp	
	2 × 150 bp		2 × 150 bp	2 × 100 bp	2 × 150 bp	
Run time	7–24 h	21–56 h	11–29 h	2 × 125 bp	< 1–3.5 days	< 3 days

bp basepairs, *Gb* gigabase pairs, *PE* paired-end sequencing, *SE* single-end sequencing

overlaps between all reads, and the reads and overlap information are laid out on a graph and consensus sequences are then inferred. This algorithm, often used with Sanger-generated data, has been widely incorporated into assembly programs such as Arachne [6], Celera Assembler [7], PCAP [8], and PHRAP [9]. Although this approach provides a cheaper and faster way of utilizing Sanger sequencing for reference genome development, with larger datasets the assemblies usually have gaps and result in unplaced scaffolds that require more effort to verify and finish. This heralded the era of draft genome assemblies and a subsequent change in standards for the quality of a reference genome.

The significantly higher data volume, shorter read lengths, and platform-specific error profiles of NGS data present challenges for algorithm developers. The higher amounts of short-read data from the next generation sequencers furthered new developments in assembly algorithms and a few overlap-layout-consensus assemblers such as Celera Assembler [7], PCAP [8], and Newbler [4] were extended from their original versions to handle both Sanger and NGS data from 454 sequencers. However, the increased usage of short read Illumina sequences for assembling large complex genomes spurred the development of the second class of assembly algorithms – those using the more efficient DBG-based approaches. The DBG approach works by first chopping reads into shorter k-mers, using those k-mers to build a graph and using the graph to infer the genome sequence. Assemblers such as ABySS [10], ALLPATHS-LG [11], and SOAPdenovo [12, 13] rely on the DBG approach for increased efficiency.

2.3 Reference Genome Project Strategies

2.3.1 Sanger-only Assemblies

Sequencing technologies have enabled the study of genomes across all spheres of life. The first genomes to be sequenced were bacterial [14, 15] and employed a whole genome shotgun approach. However, at the time, larger genomes were not considered good candidates for this approach. Consequently, a hierarchical shotgun strategy was developed for the first large genomes, including the generation of the first plant reference genome for the model plant *Arabidopsis thaliana*. The Arabidopsis Genome Initiative (AGI), an international consortium, generated comprehensive BAC libraries and used the BAC end-sequences and fingerprints of individual BAC clones to create a physical map. A minimum tiling path of BAC clones along each chromosome was identified and the selected BACs were then individually shotgun-sequenced by consortium members and assembled using assemblers such as the TIGR Assembler [16] to produce assembled contigs. The BAC ends were later used to link contigs into scaffolds and the genetic map served as a foundation for integrating assembled scaffolds into chromosomes [17].

The initial strategies for reference genomes relied predominantly on Sanger sequencing and continued to make advancements through automation or

incorporating improved methodologies. For instance, the rice genome sequences were assembled using PHRED and PHRAP software packages or the TIGR Assembler with the finishing step incorporating some automated and manual improvements and sequence gaps resolved by full sequencing of gap-bridge clones, PCR fragments, or direct sequencing of BACs [18]. The maize genome also relied on the hierarchical approach and Sanger sequencing while utilizing optical mapping to order and orient contigs into chromosomes [19].

The generation of the soybean reference genome [20] used the whole genome shotgun strategy – first used in the early bacterial genomes in 1995, and later adapted for the Celera human genome and many other mammalian genomes. The basic WGS strategy involves randomly shearing the genome and sequencing the fragments from this WGS library. The modified approach for larger genomes generates sequence libraries from multiple-sized fragments. For soybean, an initial WGS library of ~1,000 bp inserts was combined with 3, 8 kb, Fosmid and BAC libraries. The soybean sequence data were assembled using Arachne [6], where an initial assembly generated from the WGS library was combined with paired end data from multiple libraries for scaffolding the contigs [20]. Subsequently, many other plant genomes have been sequenced with this approach [21–25].

2.3.2 NGS Technologies for Reference Genome Generation

With the advent of cheaper and high-throughput NGS technologies, Sanger sequencing was soon relegated to the back seat for sequencing needs. 454 and Illumina platforms that could generate several megabases of sequence data in a short time, opened up genome projects to researchers outside of the large genome centers. Although the newer technologies produced shorter read lengths (32–500 bp), and thus presented assembly challenges, the higher throughputs, lower costs, and faster data turnaround made them hard to resist, and soon there was a surge in reference genomes from plant species, albeit with lower quality than Sanger genomes. NGS has been applied to more genomes as the cost of NGS dropped quickly (Fig. 1). About 73% of first 50 plant genomes published are on crop species and most of them include NGS as part of sequencing [26].

2.3.3 Hybrid Sanger-NGS Assemblies

Although many genome projects started to rely on NGS for generating assembled reference genomes, the contiguity from NGS-only assemblies was far shorter than those from Sanger sequencing. Thus, strategies to sequence large complex crop genomes began to rely on a combination of Illumina, Roche 454 and Sanger platforms to balance the cost and contiguity of assemblies. For example, the genome of oil seed rape, *Brassica napus*, was sequenced using a combination of multiple platforms: 21.2× coverage from GS FLX Titanium sequencing (reads of 450 bp average size), 0.1× Sanger BAC ends (reads of 650 bp average size), and 53.9× Illumina HiSeq sequencing (reads of 100 bp) [27]. The 454 sequencing included

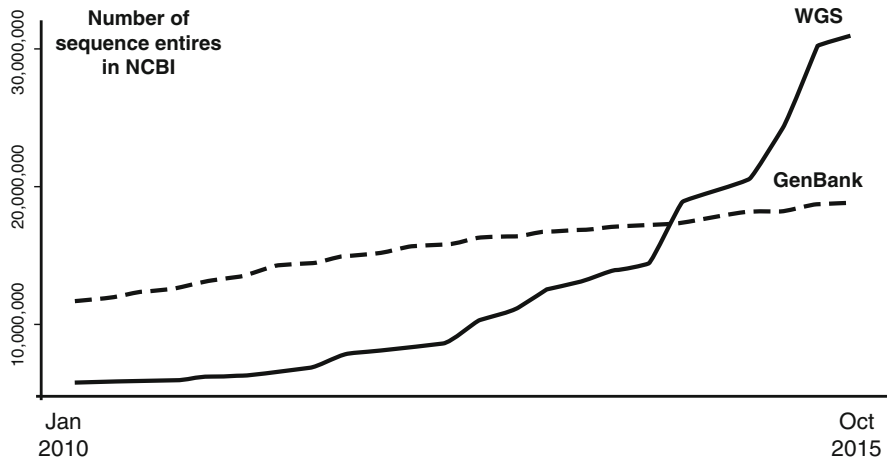


Fig. 1 Decreasing sequencing cost and increasing read length driven by introductions of new technologies. Data sources: Illumina (www.illumina.com), PacBio (www.pacb.com), Oxford Nanopore (www.nanoporetech.com)

regular 8 and 20 kb libraries and Sanger-based BAC ends were from a BAC library of 139 kb average insert size. The longer reads were assembled using Newbler to generate an initial assembly and Illumina reads were used for final error correction and gap filling with the construction of final pseudomolecules facilitated with genetic maps. A similar strategy of combining benefits from multiple technologies was used to generate the reference genomes of tomato, cassava, and African rice [28–30]. As more NGS sequences were used, a drop in assembly quality is generally seen compared to the genomes sequenced using Sanger method.

2.3.4 NGS-only Assemblies

With continuous improvements in the Illumina platform and assembly algorithms, NGS-only genomes have increased in number. The Illumina platform was used to generate a chromosome-based draft sequence of the hexaploid bread wheat [31]. High depth of Illumina sequences was also added to the *B. rapa* genome [32], the diploid [33], and allopolyploid cultivated [34, 35] cotton. Due to the repetitive nature of crop genomes, the contiguity is much lower than that from Sanger sequencing. Although the hierarchical approach algorithmically has advantages over WGS approaches, the overall process of generating a BAC library, physical map, and MTP are very labor and time intensive, making these projects very expensive and time consuming.

2.4 *Resequencing Strategies*

The availability of high-quality reference genome sequences combined with higher throughput and lower cost of sequencing is making it possible to comprehensively understand diversity within a species by generating sequence from many accessions. Whole genome resequencing is being effectively utilized to understand crop diversity and create genomic resources to enable crop improvement across a wide range of crops. This approach generates low coverage (usually $2\times$ to $10\times$) genome sequence data from accessions of interest and compares the sequences against a reference genome to detect various kinds of variation – single nucleotide polymorphisms (SNPs), insertion-deletions (InDels), presence-absence variants (PAVs), copy number variations (CNVs), and other structural variants – to understand the genetic diversity of a crop species. In plants, the 1,001 genome project in *Arabidopsis* [36] demonstrated the value of resequencing to enhance understanding of a species and soon several large-scale resequencing projects were initiated in crop plants like rice [37], maize [38, 39], soybean [40], and sorghum [41]. These resequencing data were able to provide unprecedented information about the variation existing within each crop species that can be utilized for improvement of these crops. Such resequencing data are now routinely used to find novel alleles for genes of interest [42–45], find the signals of domestication, provide background data to build genomic selection models, and form the basis for generation of tailored populations such as multi-parent advanced generation inter-cross (MAGIC) and nested association mapping (NAM) populations. Many of these applications are discussed in detail later in this volume.

Sequencing several accessions from a crop has demonstrated the presence of extensive structural variations within crop species [37, 38] leading to the recognition of the importance of generating multiple *de novo* assembled genomes (e.g., soybean, rice) [34, 35, 46]. Although high-throughput NGS technologies have shown advantages in generating variants and draft assemblies at low cost, the incompleteness of these assemblies and their reliance on a single existing reference genome makes it challenging to comprehensively identify structural variations.

In 2015, sequence entries archived in NCBI showed an interesting pattern: the number of entries for WGS surpassed the general sequence entries submitted to GenBank (Fig. 1). The dramatic increase of WGS data has been a result of re-sequencing driven by ever-decreasing sequencing cost (Fig. 2). Biologists have been using the resequencing approach for across a wide range of species and for varied research goals. For all, the ability to sequence across multiple individuals is a powerful approach, albeit with logistical challenges.

2.5 *Data Management and Visualization*

When the first plant genome became available, efforts in data management and visualization were primarily focused on making the sequence data and the corresponding annotations available to a broader scientific community and enabling

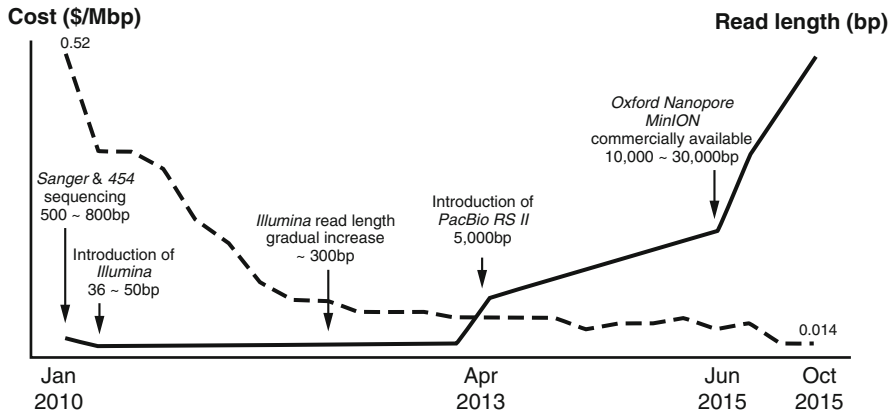


Fig. 2 Submitted sequence entries for GenBank and WGS archived in NCBI. Data source: NCBI (<http://www.ncbi.nlm.nih.gov/genbank/statistics/>)

the use of genome sequences to address specific research questions. With the large influx of genome sequencing data from NGS technologies, tools for data analysis, storage, and management soon became a critical need of the scientific research community. Initial developments centered on developing data standards and guidelines so that data could be easily shared and accessed.

2.5.1 Variant Data Standards

The human 1,000 Genomes Project led the way and provided invaluable insights into genetic variants in humans, as well as established some of the early standards to manage and analyze large-scale variant data that soon became the standard for later large-scale studies in all organisms [47–49]. New file formats to compress and store sequence alignment data and tools that could manipulate these file formats quickly came into existence and widely spread within the bioinformatics community [50]. The 1,000 Genomes Project created the Variant Call Format – a format that has become the standard for managing and manipulating variant data obtained by comparing re-sequencing data to reference genomes [51]. The initial development of VCFtools and the more recent vcfR and PyVCF tools enabled scientists using three major programming languages used in the bioinformatics community to embrace VCF as the system to manage and analyze variants [51, 52]. These tools in combination with SnpEff, a tool for annotating functional impacts of SNPs, provide a toolkit to utilize variant data in the pursuit of answers to deeper scientific questions [53].

2.5.2 Variant Data Management Systems

While the VCF file is fairly simple, it can contain essentially complete information about individual variants. However, it is not a user-friendly format for querying as VCF files can easily contain millions of variants, and be 10s or 100s of Gigabytes in size. Solutions employing relational database or indexing schemes are needed to extract information from VCF quickly and efficiently with complex query structures.

One solution to the variant storage and query problem is to utilize relational database systems such as MySQL. One example of this is the maize HapMap project [38]: all variants generated in this project are imported into “Ensembl Variations,” in which each variant, SNP or InDel, is stored as an entry in the relational database and contains several attributes linking it to other relevant information (Fig. 3). A user can explore the frequency and genome context, as well as linkage information of variants using the data schema of Ensembl.

The Ensembl MySQL solution is intended for the most widely used genomes that have Gold Standard quality assemblies, extensive annotations, and functional studies. It may not work with the majority of re-sequencing projects, as these projects are often focused on less well studied species whose data do not meet the high quality standards of Ensembl. For such projects, VCF is still the best choice for data retention and downstream analyses, but there are some alternate solutions that do not rely on relational databases. For example, genome browsers such as JBrowse render compressed and indexed VCF to visualize information [54]. The “focused” nature of a genome browser takes advantage of the indices to show only variants in selected genomic intervals.


In many cases, the primary piece of information needed is the impact of the variant, i.e., the functional annotation of the variant: is it in a coding region or non-coding region, is it a synonymous or non-synonymous change, etc. For this purpose, there are a number of solutions [53, 55, 56]. For example, SnpEff is a suite of tools for genetic variant annotation and effect prediction. A primary advantage of SnpEff is the 38,000 genomes supported out-of-the-box, so users can leverage prior annotation efforts of the community. SnpEff also supports VCF files generated by major variant calling pipelines such as SAMtools and GATK and appends the annotation results to the VCF files. The VCF-in and VCF-out workflow for SnpEff enables users to apply existing tools for manipulating VCF files and allows SnpEff to be tightly integrated into analysis pipelines without too much additional effort. Another SNP annotation tool with comparable gene annotation databases is Ensembl’s Variant Effect Predictor (VEP). Unlike SnpEff, VEP does not generate VCF files but a unique plain text, closely tied to the unique relational database of Ensembl. By taking advantage of the rich infrastructure of Ensembl’s web front end, VEP provides a more user-friendly point-and-click web interface for variant annotation.

Login/Register
Search Ensembl Plants...

HMIMER | BLAST | BioMart | Tools | Downloads | Documentation | Website help


Search: **All species** for

e.g. Carboxy* or chx28




Popular genomes


Arabidopsis thaliana
TAIR10




Oryza sativa Japonica
RGSP-1.0




Triticum aestivum
TAOv1



Hordeum vulgare
ASH2268v1



Zea mays
AGPv4



Physcomitrella patens
ASM213v1

★ [Lot in to customize this list](#)

All genomes

-- Select a species --

[View full list of all Ensembl Plants species](#)

New Bread Wheat Genome Assembly

A new genome assembly of *Triticum aestivum* cv. Chinese Spring is now available in Ensembl Plants. The assembly (TAOv1) and its accompanying annotation was produced by the Euzhan Institute, formerly The Centre for Genome Analysis (TGAC), as part of the [Triticum Genomics for Sustainable Agriculture](#) project.


The assembly has a scaffold N50 of 88 Kbp and a total length of 13.4 Gbp in contigs greater than 500 bp ([read more](#)). The gene model annotation consists of 217,907 loci and 273,739 transcripts. A total of 104,06 protein coding genes (154,798 transcripts) and 10,156 long ncRNAs have been annotated with high confidence ([read more](#)). Approximately 99,000 genes (99% of the total) annotated on the previous IWGSC CSS assembly (MIPS) have been mapped to the new assembly.

The Avicrom 35k and 820k SNP marker sets have been provided by [CerealsDB](#) and located on the new assembly ([read more](#)).

Ensembl Plants Archive Site


Alongside release 32 we have launched a new [archive site](#), where we will keep selected previous releases of Ensembl Plants publicly available. The first release available on the archive site is release 31, and includes the previous assemblies for wheat and maize.

Ensembl Plants is developed in coordination with other plant genomics and bioinformatics groups via the EBIs role in the [transPLANT](#) consortium. The [transPLANT](#) project is funded by the [European Commission](#) within its [7th Framework Programme](#), under the thematic area "Infrastructure", contract number [2634106](#).



Part of the
transPLANT
European Plant
Genomics Infrastructure

Wheat genomics resources are developed as part of our involvement in the consortium [Triticum Genomics For Sustainable Agriculture](#). Barley genomics resources are funded through the [UK Barley Genome Sequencing Project](#). Both projects are funded by the BBSRC.



BBSRC
Bioscience for the future

What's New in Release 32

- New genomes
 - [Beta vulgaris](#)
 - [Brassica napus](#)
 - [Triticum pratense](#)

Did you know...?
 You can search the [Track](#)
[Hub Registry](#) to find more
[transcript models](#), [SNPs](#), & [CDS](#)

Fig. 3 Ensembl variations: explore one variant at a time

2.5.3 Visualization of Variant Data

Many layers of information are stored in the linear string of four nucleotides that make up a genome – from single nucleotides, codons, exons and genes to regulatory units, chromatin structure, and chromosome conformation. Visualization of re-sequencing results at many different levels is a crucial component of such projects. Generally, there are two approaches to visualizing data (primarily reads and/or VCF files) from re-sequencing projects: one is the dedicated application on a desktop or laptop computer; the other is by utilizing an Application Programming Interface (API) for existing web-based genome browsers to work with short-read mapping and variant calling results. Given the amount of data from re-sequencing projects, the key to achieving performance is to create the ability to access only the reads or variants needed for the specific slice of genome that is being viewed.

The champion of read-centric visualization tools is Integrative Genomics Viewer (IGV) [57]. In addition to providing a large number of ready-to-use genomes and annotations, IGV has the best support for visualizing almost every detail of read-mapping information, including the very important but largely overlooked CIGAR string [50]. It also provides a read coloring system that helps users spot split reads and read pairs with abnormal insert sizes between the mates – crucial for the exploration of structural variations. IGV has also gone beyond a standalone desktop application and supports the access of data files from distributed sources via the HTTP protocol. Tablet is another desktop solution for read visualization that stands out from the crowd with its great usability and interface. Tablet works extremely well in terms of zooming in and out, as well as views at different levels in one screen (Fig. 4).

With the need to visualize the large amount of re-sequencing data, the traditional feature-based genome browsers are playing a catch-up game. Genome browsers, such as UCSC browser and GBrowse, have been the data hub and integrator for feature-based data, i.e., genomic data based on genomic intervals for many years [58, 59]. The feature-based data are rich, detailed, but small in size, so the traditional genome browsers have been optimized to primarily handle large numbers of tracks of small sizes. NGS data from re-sequencing projects present the opposite challenge – read alignments files are very big, but information for each read is minimal. Because UCSC and GBrowse both use relational databases in the backend, they had to create database adaptors to handle read alignments, which turned out to be inefficient and awkward, especially when the alignment files are large in size. Subsequently, many new genome browsers have been developed with optimized functionalities for visualizing short reads. The best examples among these are JBrowse, a generic genome browser from GMOD, and Savant Genome Browser, a short-read browser optimized for human genome and medical and diagnostic purposes [54, 60]. Both genome browsers abandoned the old relational database architecture and embraced read alignment formats directly, so they read the alignments and render the reads on-the-fly. Coupled with various indexing schemes, they

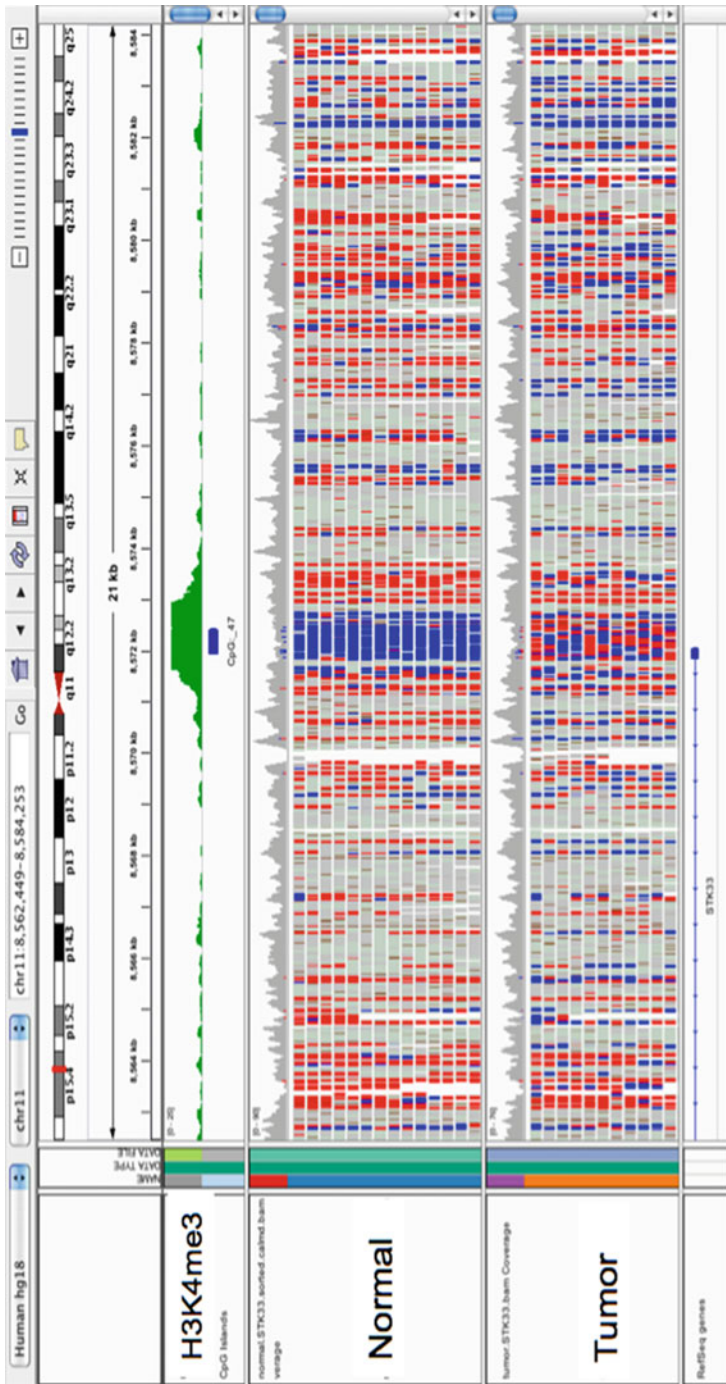


Fig. 4 IGV and Tablet visualization of reads and read alignments. Data sources: Broad Institute (software.broadinstitute.org), Tablet (ics.hutton.ac.uk/tablet/)

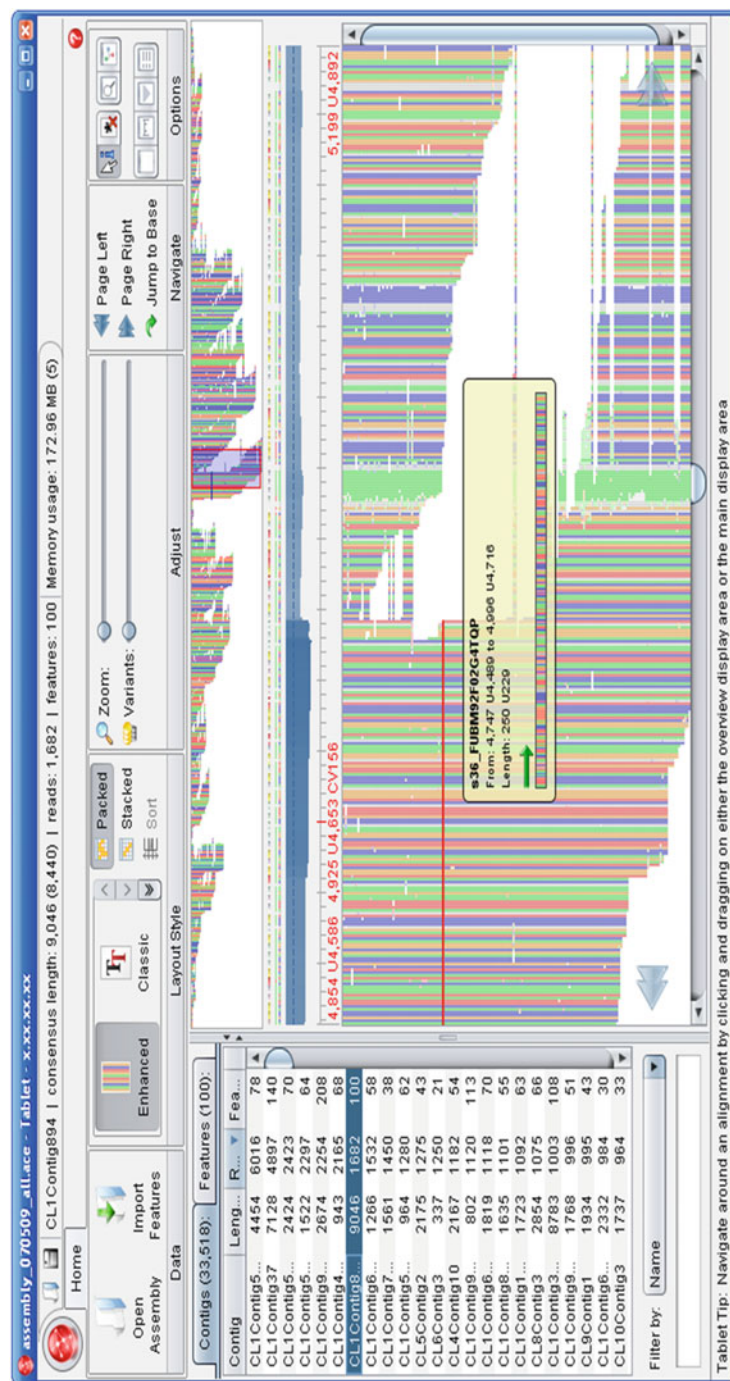


Fig. 4 (continued)

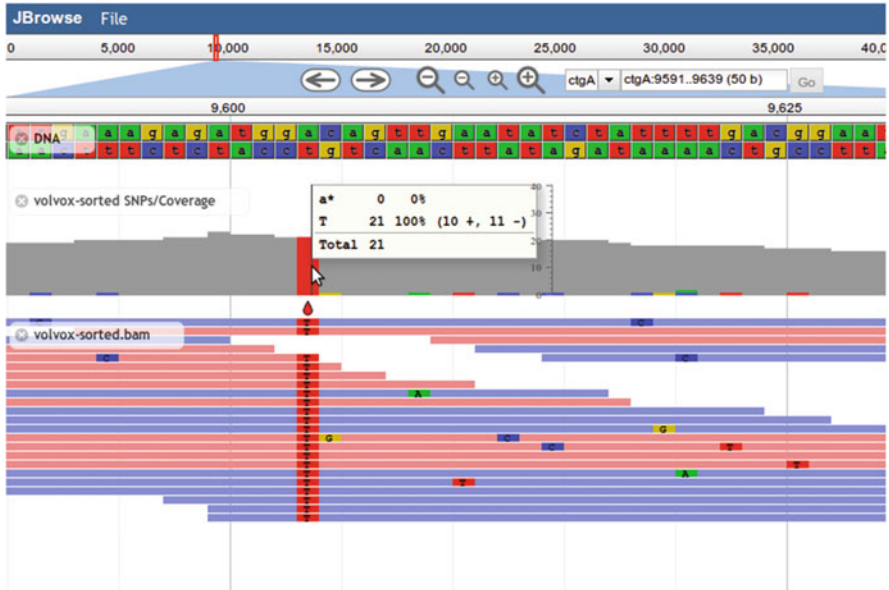


Fig. 5 Rendered reads and SNP in a JBrowse view, adapted from Ref. [54]

provide intuitive navigating functions to explore read mapping and variant calling results (Fig. 5).

The advantage of generic genome browsers over a specialized short reads viewer is that it is very easy to incorporate feature-based genomic data and much more into a holistic genomic view. Instead of adding short read functionalities as an afterthought, this new generation of genome browsers puts short reads in the center and builds genomic resources around them. Moreover, a lot of the new genome browsers have started to utilize cloud platforms to store and manage sequencing data, majority of them re-sequencing data (<https://cloud.google.com/genomics/>, <https://aws.amazon.com/>). This should not be surprising since the “big” nature of re-sequencing data fits nicely to the concept of “Big Data” advocated by cloud technologies.

3 Trends, Advanced Technologies, and Strategies

3.1 Sequencing Technologies

The second-generation sequencing technologies such as Illumina are very useful for resequencing studies to understand the variability of a crop species. However, due to their short reads, it is challenging to generate finished quality reference genomes. The recent emergence of long-read sequencing technologies such as PacBio (<http://www.pacb.com/>) and Oxford Nanopore (<https://www.nanoporetech.com/>), and technologies that focus on providing long-range genomic linking information such

as Dovetail Genomics (<http://dovetailgenomics.com/>), 10× Genomics (<http://www.10xgenomics.com/>) and BioNano Genomics (<http://www.bionano-genomics.com/>) are making it feasible to generate good quality reference genomes faster and cheaper.

PacBio single-molecule real-time (SMRT) sequencing captures the sequence information during the replication process of a DNA molecule that is tracked in a zero-mode waveguide (ZMW) on a SMRT cell. The DNA molecule is circularized by adding the adapters on both ends and diffused into a ZMW with DNA polymerase immobilized at the bottom. Four fluorescent bases are flowed through the SMRT cell and a distinct light pulse is produced for each base that is recorded as a movie. The movie can then be analyzed to extract DNA sequence. PacBio can produce reads in the average range of 20 kb and is being routinely used to finish microbial genomes [61]. Until recently, it has been very expensive to use PacBio data alone for a large crop genome, and thus many hybrid strategies have been deployed that combine PacBio sequences with other short read data to improve genome assemblies [62], and new algorithmic strategies are being developed to better utilize these long reads in assembly processes both in hybrid strategies and alone [63–66]. The new SEQUEL system from PacBio can deliver up to 50 Gb sequences for a few thousand dollars at an average read length of ~20 kb and consensus accuracy >99.999%, making it an attractive option for crop reference genomes. For example, the genome of adzuki bean (*Vigna angularis*) was assembled using SMRT sequencing technology and the PacBio assembly produced 100 times longer contigs with 100 times fewer gaps compared to the NGS-based assemblies [67]. Efforts are currently underway to improve the B73 reference genome of maize and build high-quality reference genomes for 23 species of rice using PacBio SMRT Sequencing and create new resources for crop improvement (<http://www.pacb.com/wp-content/uploads/agi-rod-wing-corelab.pdf>).

Oxford Nanopore (ONT) sequencing is the latest long-read sequencing technology that offers a lot of promise for generating de novo assemblies of complex plant genomes. This technology passes a long DNA molecule through a charged protein nanopore and measures the changes in current as the molecule passes through the nanopore. The changes in current or “squiggleplot” are then input into a basecaller to produce DNA sequence information. ONT is very promising technology with reads as long as 150 kb having been reported by early users, although average read lengths are much lower. The technology is deployed in two forms – a small mobile sequencer, the minION, which is approximately the size of a stapler that has flowcells with 512 nanopores, and a much larger format called the promethION, which can house 48 flowcells, each with 3,000 nanopores. MinION has been commercially available since May 2015 and has been applied to the rapid identification of viral pathogens [42, 68], 16S sequencing [69], and haplotype sequencing [70]. At the time of writing, nearly 50 publications have used or developed tools for the ONT platform. As the accuracy and throughput continue to improve, de novo sequencing of large complex crop genomes will become practical soon.

The parallel development of several long-range sequencing technologies from Dovetail Genomics and 10× Genomics, or long-range mapping technologies from BioNano Genomics, can provide the contiguity information in a genome. The long-range information when complemented with sequences from long-read single

molecule technologies can deliver high quality assemblies with fewer gaps and megabase-long contigs for complex plant genomes without the need to construct traditional physical and genetic maps. In view of the significant structural variation in crop species, these new technologies can redefine our understanding of genomes and help pinpoint the underlying genetics of complex traits.

3.2 Assembly Strategies/Technologies

Developers of assembly algorithms are focusing on developing methods that will enable the seamless integration of long-read and long-range data into the assembly process. The long-read technologies in their initial growth cycles have higher error rates, and algorithms must take these into account. Various software tools have been developed to handle multiple scenarios involving longer reads. PBJelly2 is effective on low coverage ($<15\times$) PacBio data and has the ability to use the long-read data to link scaffolds and close gaps for existing short-read assemblies [65]. Tools such as ECTools [66], SPAdes [63], and PBcR [64] can handle $20\text{--}30\times$ coverage PacBio data either in combination with Illumina reads or on their own. If more than $50\times$ coverage is available, the PacBio sequences can be assembled de novo without short-read sequences using packages such as HGAP [71] and Canu [64]. In some of the most recent algorithms, with $>30\times$ coverage of PacBio sequences, an overlap-layout-consensus approach can be used to assemble corrected sequences. A final polishing step is used to correct the errors in the consensus with raw PacBio sequences, which can improve the consensus accuracy to 99.999%. Similar assembly and error correction strategies can be applied to ONT data. As each platform continues to improve accuracy towards a 1% error rate, error correction will not be necessary.

3.3 Genome Project Strategies

The development of new third-generation sequencing technologies is leading to a trend of combining these data with second generation technologies in genome sequencing projects. Due to a relatively lower throughput and higher cost (per Gb) of the long-read technologies, current genome sequencing strategies typically combine lower coverage long-read/long-range data to with higher coverage of short read data to improve the qualities of genome assemblies especially for large, complex crop genomes. Ultimately, the selection of a strategy is driven by what can provide the highest quality for a given cost combined with the perception of what is an acceptable cost for a genome project. As the technologies continue to develop further, error rates and cost are expected to drop, which will change both what is possible and the perception of what is reasonable.

3.4 *Resequencing Strategies*

Short-read technologies have been heavily used by projects to generate understanding of the variation within a crop species. However, since they rely on a reference genome, these projects have had limitations in identifying large structural variations among accessions within a species. Crop species like maize and soybean have been shown to have large variation in their genome content between lines – almost to the extent of 30% [72]. Resequencing that relies on a single reference genome is not able to adequately capture the full extent of these variations. As long-read technologies continue to improve and drop in price, we expect projects to generate de novo reference assemblies for multiple lines within a species – perhaps for all lines within a species if the price drops far enough. These types of data will enable us to better characterize and catalog the variation within a species.

3.5 *Data Management, Visualization, and Storage*

The availability of multiple reference genomes for crops, the widespread re-sequencing efforts, and the resulting variant data are constantly pushing the limits of VCF files, as well as the tools and infrastructure for handling them. For example, one VCF file containing millions of SNP/InDels from hundreds of thousands of samples could not be effectively managed by any of the tools previously described. Further evolution of variant storage is needed. One recent advance is BGT, a flexible genotype query tool that works with large scale multi-sample VCF files [73]. The key to these tools is to generate indices of variant genotypes that can be harnessed for rapid retrieval in a flexible manner – whether it be for a subset of genomic locations, or a subset of samples or any combination of both. Moreover, BGT supports encoded phenotypes associated with the samples, which creates opportunities to slice and group the samples based on phenotypes, a convenient way to conduct local and small-scale association studies.

Meshing the concepts contained within VCFtools, PyVCF, vcfR, and BGT, there are unlimited possibilities to create new tools to extract the information in VCF and utilize the information in plant genetics and crop breeding. To date, information from VCF files has been queried, summarized, and manipulated for GWAS, LD analyses, small-scale association studies, as well as variant data dissemination through various data visualization frameworks. We expect this trend to continue with ever more sophisticated data manipulation tools and approaches.

3.6 *Beyond Individual Variants: Alleles, Haplotypes, LD Blocks, and Pan-Genomes*

Interpretation of the biological meaning of information contained within sequence data is not the exclusive domain of bioinformaticians. It takes collaborations between biologists of all kinds, which is particularly true as the complexity of the

data and data structure increases. Information stored in genomes comes at several levels, and the bulk of what we have discussed in this chapter is focused on the discovery and description of individual variants. But this just scratches the surface of the full impact of genomic diversity. There are relationships between variants that can be identified, stored, and visualized.

As an example, consider a gene as a unit within which combinations of multiple variants create the variant of that gene that may have a phenotypic impact. Each combination of variants that form a single version of that gene can be considered as a group to define a single allele of that gene. This requires a more complex form of annotation and visual representation than is found in current genome browsers. Genetic linkage beyond gene boundaries forms yet another higher level of information, indicating longer segments that are segregating and propagating in the real world, called haplotypes. Stretches of such haplotype segments form Linkage Disequilibrium (LD) blocks, in which most native traits harbor. Above the LD blocks and haplotypes, there are chromatin structure and chromosome conformations. Characterization of these even higher levels lies beyond the scope of simple re-sequencing and requires other technologies such as methylation profiling by bisulfate sequencing and Hi-C profiles [74, 75].

Strategies to explore and exploit the alleles, haplotypes and LD blocks present within crop plants are active areas of development, both in academia and commercial breeding contexts. The common goal is understanding the genetic structure of populations at a more sophisticated level than individual variants, which can then enable a mix-and-match approach to the traits required in breeding and efficient and accurate trait characterizations at the molecular level.

One other recent trend is a gradual shift away from the concept of a single reference used as a basis for all future studies of variation within that species. This is happening for a number of reasons: the reference accession is often the most “well-behaved” accession for research, rather than the best representative of the species; our perceptions of what is possible have changed along with sequencing costs; data from early re-sequencing studies showed that it is unlikely that any single reference could represent a species, even with a robust way of describing variants. The concept of the composite genome of a species, rather than an accession, is known as the pan-genome [12, 13, 76]. Depending on the evolutionary history and divergence among the individuals, pan-genomes can be very simple in closely related individuals, or very complex in groups with tremendous genetic diversity. In the latter case, a pan-genome captures the true nature of genetic variations in a way that a single reference could not, because the variation is far beyond single nucleotide changes. Maize is an example of such a species with extremely rich diversity among varieties and accessions. Variants discovered by comparing re-sequencing data to the B73 reference missed significant fractions of the true variation [38].

Unfortunately, although a simple concept, representing pan-genomes in a file format is not a simple task. The definition and implementation of pan-genomes is still in its infancy. One promising idea is to present the genomes in a graph with genomes represented by a “path” in a hypothetical space and variations represented by “bubbles” that bulge on the sides of the paths [77] (<https://www.technologyreview.com>).

com/s/537916/rebooting-the-human-genome/). The graph theory supporting the pan-genomes, by its nature, is capable of capturing and presenting information at multiple levels from single nucleotides to large linkage blocks. This will lend itself naturally to representing haplotype LD blocks useful for exploitation of variation in breeding programs. As more and more data are added, the pan-genome concept implemented with effective visualizations and query tools are going to be essential in order to gain biological insights from these incredibly valuable datasets.

4 Conclusion and Outlook

As described in this chapter, technology advances over the last decade have been tremendous and have provided great benefits across the spectrum of biological research. The low cost and ease with which sequence data can be generated has led to larger and larger experiments being imagined by more and more individual scientists. No longer do genome-wide studies require international consortia. To coin an overused phrase – we are seeing the democratization of genomics. While the challenges for individual experiments may be shrinking, the challenges for community-wide management of these enormous stockpiles of sequence data are expanding. To truly enable the next wave of genomics-enabled research, the next advances will need to be not in sequence technology but in the management, access, and analysis of exabytes of data (1 exabyte = 1 million terabytes). Just as success in automated sequencing required biologists to recruit the skills of engineers, physicists, and chemists, success in this next phase will require us to embrace those skilled in computer and data science, and computational and network infrastructure. Only one thing is certain, there will be more data tomorrow than yesterday – which is fortunate, because in science, tomorrow always unveils more questions for us to answer.

References

1. Sanger F et al (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74(12):5463–5467
2. Mardis ER (2013) Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto, Calif)* 6:287–303
3. Goodwin S et al (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6):333–351
4. Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380
5. Li Z et al (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* 11(1):25–37
6. Jaffe DB et al (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13(1):91–96
7. Myers EW et al (2000) A whole-genome assembly of *Drosophila*. *Science* 287(5461):2196–2204
8. Huang XQ et al (2003) PCAP: a whole-genome assembly program. *Genome Res* 13(9):2164–2170

9. Ewing B et al (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8(3):175–185
10. Simpson JT et al (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19(6):1117–1123
11. Gnerre S et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108(4):1513–1518
12. Li R et al (2010) Building the sequence map of the human pan-genome. *Nat Biotechnol* 28(1):57–63
13. Li RQ et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265–272
14. Fleischmann RD et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496–512
15. Fraser CM et al (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270(5235):397–403
16. Sutton GG et al (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci Technol* 1(1):9–19
17. Hamilton JP, Buell CR (2012) Advances in plant genome sequencing. *Plant J* 70(1):177–190
18. Matsumoto T et al (2005) The map-based sequence of the rice genome. *Nature* 436(7052):793–800
19. Schnable PS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956):1112–1115
20. Schmutz J et al (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463(7278):178–183
21. Goff SA et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica). *Science* 296(5565):92–100
22. Ming R et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452(7190):991–U997
23. Paterson AH et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457(7229):551–556
24. Vogel JP et al (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463(7282):763–768
25. Yu J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp indica). *Science* 296(5565):79–92
26. Michael TP, Jackson S (2013) The first 50 plant genomes. *Plant Genome* 6(2)
27. Chalhouh B et al (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345(6199):950–953
28. Prochnik S et al (2012) The cassava genome: current progress, future directions. *Trop Plant Biol* 5(1):88–94
29. Sato S et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635–641
30. Wang M et al (2014) The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat Genet* 46(9):982–988
31. International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345(6194):1251788
32. Wang XW et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet* 43(10):1035–U1157
33. Wang K et al (2012) The draft genome of a diploid cotton *Gossypium raimondii*. *Nat Genet* 44(10):1098–1103
34. Li FG et al (2014) Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat Genet* 46(6):567–572
35. Li YH et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* 32(10):1045–1052

36. Cao J et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* 43(10):956–963
37. Xu X et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 30(1):105–111
38. Chia JM et al (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* 44(7):803–807
39. Jiao Y et al (2012) Genome-wide genetic changes during modern breeding of maize. *Nat Genet* 44(7):812–815
40. Patil G et al (2016) Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. *Sci Rep* 6:19199
41. Mace ES et al (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat Commun* 4:2320
42. Bradley P et al (2015) Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun* 6:10063
43. Brozynska M et al (2016) Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol J* 14(4):1070–1085
44. Leung H et al (2015) Allele mining and enhanced genetic recombination for rice breeding. *Rice (N Y)* 8(1):34
45. Yang J et al (2015) Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. *Plant J* 84(3):587–596
46. Schatz MC 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 Biol* 15(11):506
47. Genomes Project Consortium et al (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073
48. Genomes Project Consortium et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65
49. Genomes Project Consortium et al (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74
50. Li H et al (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079
51. Danecek P et al (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158
52. Knaus BJ, Grunwald NJ (2016) VCFR: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour* 17(1):44–53
53. Cingolani P et al (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 (Austin)* 6(2):80–92
54. Skinner ME et al (2009) JBrowse: a next-generation genome browser. *Genome Res* 19(9):1630–1638
55. McLaren W et al (2016) The ensembl variant effect predictor. *Genome Biol* 17(1):122
56. Wang K et al (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164
57. Robinson JT et al (2011) Integrative genomics viewer. *Nat Biotechnol* 29(1):24–26
58. Donlin MJ (2009) Using the generic genome browser (GBrowse). *Curr Protoc Bioinformatics* Chapter 9: Unit 9.9
59. Kent WJ et al (2002) The human genome browser at UCSC. *Genome Res* 12(6):996–1006
60. Fiume M et al (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics* 26(16):1938–1944
61. Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* 23:110–120
62. Ming R et al (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* 47(12):1435–1442

63. Bankevich A et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19(5):455–477
64. Berlin K et al (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing (vol 33, pg 623, 2015). *Nat Biotechnol* 33(10):1109–1109
65. English AC et al (2012) Mind the gap: upgrading genomes with Pacific biosciences RS long-read sequencing technology. *PLoS One* 7(11):e47768
66. Koren S et al (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* 30(7):692–700
67. Sakai H et al (2015) The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Sci Rep* 5:16780
68. Quick J et al (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530(7589):228–232
69. Benitez-Paez A et al (2016) Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION (TM) portable nanopore sequencer. *Gigascience* 5:4
70. Ammar R et al (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Res* 4:17
71. Chin CS et al (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10(6):563
72. Gore MA et al (2009) A first-generation haplotype map of maize. *Science* 326(5956):1115–1117
73. Li H (2016) BGT: efficient and flexible genotype query across many samples. *Bioinformatics* 32(4):590–592
74. Belton JM et al (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58(3):268–276
75. van Berkum NL et al (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* 39
76. Hirsch CN et al (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 26(1):121–135
77. Lu F et al (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun* 6:6914

Revolution in Genotyping Platforms for Crop Improvement



Armin Scheben, Jacqueline Batley, and David Edwards

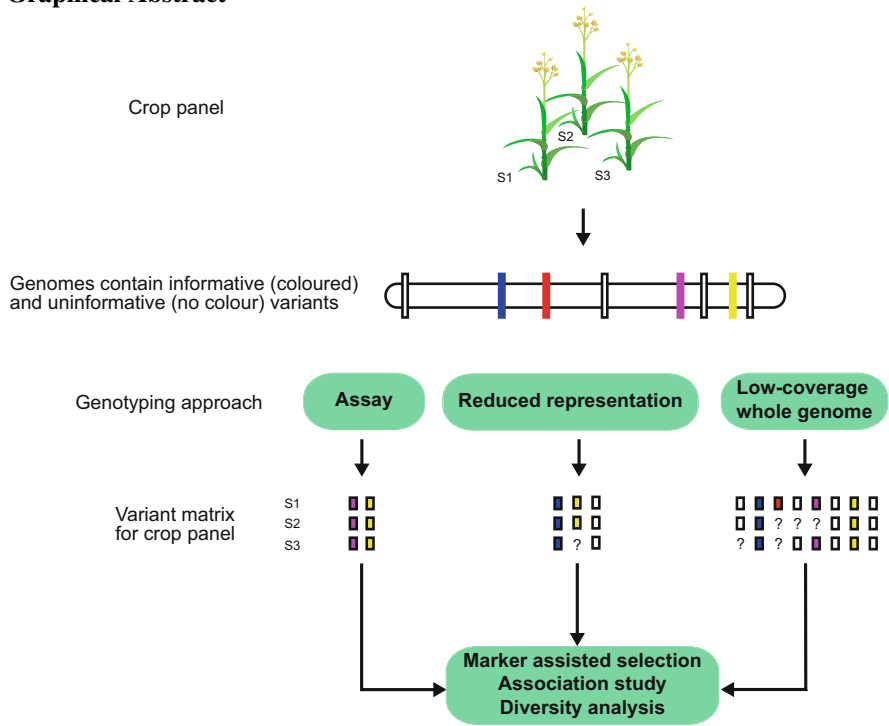
Abstract In the past decade, the application of high-throughput sequencing to crop genotyping has given rise to novel platforms capable of genotyping tens of thousands of genome-wide DNA markers. Coupled with the decreasing costs of sequencing, this rapid increase in markers allows accelerated and highly accurate genotyping of entire crop populations and diversity sets using single nucleotide polymorphisms (SNPs). These revolutionary advances accelerate crop improvement by facilitating a more precise connection of phenotype to genotype through association studies, linkage mapping and diversity analysis. The platforms driving the advances in genotyping are array technologies and genotyping by sequencing (GBS) methods, which include both low-coverage whole genome resequencing (skim sequencing) and reduced representation sequencing (RRS) approaches. Here, we outline and compare these genotyping platforms and provide a perspective on the promising future of crop genotyping. While SNP arrays provide high quality, simple handling, and unchallenging analysis, the lower cost of RRS and the greater data volume produced by skim sequencing suggest that use of GBS will become more prevalent in crop genomics as sequencing costs decrease and data analysis becomes more streamlined.

A. Scheben
School of Biological Sciences, University of Western Australia, Crawley, WA, Australia

J. Batley and D. Edwards (✉)
School of Biological Sciences, University of Western Australia, Crawley, WA, Australia

Institute of Agriculture, University of Western Australia, Crawley, WA, Australia
e-mail: Dave.Edwards@uwa.edu.au

Graphical Abstract



Keywords Breeding, Crops, Genotyping-by-sequencing, Restriction site associated DNA sequencing, Single nucleotide polymorphism (SNP), Whole genome sequencing

Contents

1 The Advent of High-Throughput Genotyping 39

2 Genotyping by Sequencing 39

 2.1 Reduced Representation Sequencing 40

 2.2 Skim Sequencing 40

3 High-Throughput SNP Assays 41

4 Comparison of Genotyping Platforms 43

5 Conclusion and Perspectives 44

References 44

1 The Advent of High-Throughput Genotyping

The study of DNA polymorphism in crops is a fundamental step in molecular breeding programs. Using molecular tools to determine the DNA polymorphisms of an individual plant, relative to other individuals or species, is known as genotyping. Genotyping has many applications including marker-assisted selection (MAS), genomic selection (GS), linking phenotypes to genes, DNA barcoding, diversity analysis, and improving genome assemblies. Although crop genotyping has been common practice since the 1990s, genotyping platforms and the types of polymorphisms used have undergone rapid changes in the past decade [1]. While earlier methods relied on random amplification of PCR fragments, patterns of DNA restriction digestion and hybridization, the advance of high-throughput sequencing technologies and sophisticated assays enables genotyping on a much greater scale and at higher resolution [2, 3]. As part of this transition, the now widespread use of single nucleotide polymorphisms (SNPs), the most frequent type of variation in the genome [4], has increased the number of markers often used from fewer than 100 to tens of thousands. Furthermore, the single-base resolution of SNPs allows better detection of markers causally linked to agronomic traits.

Past genotyping efforts have had relatively little impact on breeding practice with regard to complex traits, because loci linked to traits of interest were not well resolved or had only weak individual effects and thus could not provide the returns required by costly MAS. High-throughput genotyping and novel bioinformatics tools [5, 6] now help resolve this issue by reducing the cost of genotyping while increasing accuracy. This revolution in genotyping heralds a more substantial impact of crop genomics on breeding programs in the future [7].

The two most common types of high-throughput genotyping platforms are commercial SNP assays and genotyping by sequencing (GBS). The most widely used SNP assays are the Illumina Infinium assay and the Affymetrix GeneChip[®], while GBS encompasses a range of methods using either reduced representation or whole genome approaches for characterizing polymorphisms in genomes. These high-throughput platforms each offer an unprecedented scale and quality of genomic data, providing a foundation for accelerated crop improvement [8, 9].

2 Genotyping by Sequencing

GBS was first used as a combination of restriction site associated DNA (RAD) techniques and high-throughput sequencing [10, 11]. This allowed the reduction of genome complexity before sequencing, reducing per-sample costs and effort required for data analysis. Reduced representation sequencing (RRS) thus became popular because it allowed the collection of data sampled from across most of the genome for hundreds of samples, while avoiding the cost of deep whole genome sequencing. A different solution implemented in rice employed low coverage whole

genome sequencing, or skim sequencing [12]. RRS has gained increasing popularity in recent years and protocols have been modified and developed further into over a dozen approaches. Skim sequencing has also been increasingly adopted in studies on crops [13–15], and numerous analytical approaches have been developed to overcome the often lower confidence of SNP calls for low coverage data [12, 13, 16, 17].

2.1 *Reduced Representation Sequencing*

The original RADseq [10, 11] follows a six-step protocol. First, genomic DNA is digested with a single restriction enzyme. Adapters with barcodes are ligated onto the digested ends to enable the sequencing of multiple samples in a single lane. After a sonication step, an adapter is ligated to the randomly sheared end. In the final steps, the library is size-selected, and RAD fragments with both adapters are PCR amplified. The original RADseq was used to develop linkage maps and conduct QTL analysis in crops as diverse as aubergine [18] and barley [19]. Elshire et al. [20] simplified the original RADseq protocol to four steps by implementing digestion and adapter ligation in a single well and eliminating random shearing and size selection steps. Their “GBS” technique is given in inverted commas here to differentiate it from the umbrella term GBS, which came into use after the method was developed. In “GBS,” barcoded adapters and common adapters with an overhang matching the restriction site are ligated onto digested fragments in a single sticky-end ligation. While original RADseq involves sequencing fragments to high coverage, the focus of “GBS” is to sequence with low coverage. This technique has been successfully used in a number of species, generating 24,186 genome-wide markers in barley [20], and 30,984 high-quality SNPs in rice [21].

Another important step in the diversification of RAD methods was the introduction of two enzymes in the double-digest RAD protocol (ddRAD) [22]. Combining a low-frequency and high-frequency cutter to digest DNA, a barcoded adapter is ligated to one restriction site and a common adapter to the other restriction site. Samples are then pooled and size-selected. Lastly, PCR is used to enrich the library and also introduce a second barcode in the form of an Illumina index, increasing multiplexing potential. Similar to this approach is two-enzyme “GBS” [23], which also uses two restriction enzymes. The ddRAD method has been employed for genetic linkage mapping in cultivated peanut [24] and for linkage disequilibrium and association analysis in *Brassica napus*, detecting two loci associated with seed oil content [25].

2.2 *Skim Sequencing*

Skim sequencing differs from RRS in the lack of complexity reduction steps before sequencing. To make genotyping large populations cost effective, sequencing is carried out at low coverage depth, typically between $1\times$ and $5\times$ or even lower [12, 13]. To

simplify data analysis, heterozygous alleles are often eliminated by sequencing recombinant inbred lines (RIL) or double haploid (DH) populations. The parental genomes and a reference sequence are commonly required for these mapping populations, though they can also be inferred using hidden Markov models [17], reducing the cost for deep sequencing of the two parents. Training the model on each individual sample refines this approach by allowing for variation in error rates [16]. This method is particularly useful in genotyping a constructed cross population, in which the parental lines are not known and parental genome sequences are not yet determined.

Skim sequencing has allowed detection of 270,820 high quality SNPs and identification of grain weight QTLs in rice [26]. Genotyping by resequencing has been applied frequently in rice, e.g., a total of 1,493,461 SNPs were identified in 150 RIL sequenced at $0.02\times$ coverage. Using recombination bins to construct a linkage map, it was possible to identify 49 QTLs, including four linked to plant height [12]. In sorghum, the same approach for 244 RILs sequenced at $\sim 0.07\times$ coverage led to the discovery of 7.76 million high-quality SNPs and, after map construction, the identification of several major QTLs for heading date and plant height [27]. Finally, skim GBS genotyping of chickpea and rapeseed identified 511,624 SNPs and 794,837 post-filtered SNPs respectively. Based on these SNPs, numerous crossovers and gene conversions in both species could be identified [13]. In a further study on chickpea using skim sequencing, 53,169 post-filtered SNPs were detected and used for QTL analysis to identify four candidate genes implicated in drought tolerance [15].

3 High-Throughput SNP Assays

High-density genotyping assays, or “SNP chips,” are a valuable resource for genomic studies in crops. The commercial SNP assays available from Illumina and Affymetrix rely on distinct technologies, but are both capable of producing highly scalable assays. Affymetrix’s hybridization arrays and Illumina’s Infinium-based arrays enable parallel genotyping of hundreds of samples for hundreds of thousands to millions of SNPs (<http://www.illumina.com>; <http://www.affymetrix.com>). Illumina’s Infinium BeadChip[®] assay utilizes beads covered with specific oligos that fit into patterned microwells and is highly scalable, as the number of SNPs genotyped can be increased with higher densities of microwells. The assays are based on a two-color single-base extension from a single hybridization probe per SNP marker [28]. The GeneChip[®] array of Affymetrix arrays, on the other hand, use photolithographic printing of oligos on an array, with the Affymetrix Axiom[®] technology based on a two-color, ligation-based assay with 30-mer probes. Currently, arrays from Affymetrix and Illumina are available for many common crop plants (Table 1). To reduce costs, these arrays are usually developed by agrigenomic consortia. Illumina currently offers a larger selection of DNA genotyping arrays, while Affymetrix has a larger selection of expression arrays. Recently an Infinium genotyping array for 90,000 gene-associated SNPs in wheat was developed and

Table 1 Crop species with commercial DNA genotyping arrays available from Affymetrix or Illumina

Affymetrix	Reference
Apple	Bianco et al. [29]
Broccoli	Vosman et al. [30]
Chickpea	Roorkiwal et al. [31]
Cotton	Rai et al. [32], Byers et al. [33]
Groundnut	Pandey et al. [34]
Lettuce	Stoffel et al. [35]
Maize	Unterseer et al. [36]
Capsicum	Hill et al. [37]
Rice	Zhao et al. [38], Yu et al. [39], Singh et al. [40]
Rose	Koning-Boucoiran et al. [41]
Strawberry	Bassil et al. [42]
Soybean	Lee et al. [43], Wang et al. [44]
Wheat	Winfield et al. [45]
Illumina	Reference
Alfalfa	Li et al. [46]
Apple	Bianco et al. [47, 48]
Barley	Comadran et al. [49], Rostoks et al. [50], Close et al. [51]
Bean	Song et al. [52]
<i>Brassica napus</i>	Snowdon and Luy [53], Edwards et al. [9], Dalton-Morgan et al. [54], Durstewitz et al. [55], Delourme et al. [56], Clarke et al. [57]
Cherry	Peace et al. [58]
Chickpea	Choudhary et al. [59], Roorkiwal et al. [60], Gaur et al. [61], Bajaj et al. [62]
Cocoa	Livingstone et al. [63]
Cotton	Hulse-Kemp et al. [64]
Cowpea	Close et al. [65], Muchero et al. [66]
Eucalyptus	Silva et al. [67]
Grape	Myles et al. [68]
Maize	Ganal et al. [69], Yan et al. [70], Rousselle et al. [71], Tian et al. [72]
Oat	Tinker et al. [73], Oliver et al. [74]
Pea	Tayeh et al. [75], Deulvot et al. [76]
Peach	Verde et al. [77]
Pepper	Ashrafi et al. [78]
Perennial ryegrass	Blackmore et al. [79], Paina et al. [80], Studer et al. [81]
<i>Pinus taeda</i>	Plomion et al. [82]
<i>Populus trichocarpa</i>	Geraldes et al. [83]
Potato	Hamilton et al. [84], Felcher et al. [85]
Rice	Chen et al. [86], Zhao et al. [87], Felcher et al. [85], Travis et al. [88], Ye et al. [89], Thomson [90]
Rye	Haseneyer et al. [91]
Sorghum	Bekele et al. [92]
Soybean	Song et al. [93], Hyten et al. [94], Akond et al. [95]

(continued)

Table 1 (continued)

Platform	Reference
Illumina	Reference
Sunflower	Bachlava et al. [96], Talukder et al. [97]
Tomato	Sim et al. [98]
Wheat	Wang et al. [99], Akhunov et al. [100], Cavanagh et al. [101]

46,977 of these SNPs used to create a genetic map [99]. In capsicum, Hill et al. [37] developed an Affymetrix GeneChip[®] array for 30,000 gene-associated SNPs, with the goal of facilitating the introgression of agronomic traits such as disease resistance into the breeding germplasm. High-throughput genotyping using arrays has also been applied to facilitate accurate germplasm identification in Brassicaceae, demonstrating the potential of the method to increase the value of germplasm collections for plant breeders [102].

4 Comparison of Genotyping Platforms

SNP arrays and the various GBS platforms differ in SNP discovery rate, evenness of sampling, cost, time, and effort required per sample. Depending on the number and type of samples and the SNP density needed, the advantages of one platform may outweigh those of the others. SNP arrays have been used extensively in crops and livestock and offer several distinct advantages over GBS. These advantages include robust allele calling with high call rates, low cost per sample when genotyping large numbers of samples, and simple data analysis. By selecting SNPs rather than sampling at random across the genome, SNP arrays may also provide substantially more power than randomly chosen SNPs such as those in skim sequencing. The use of fewer redundant SNPs than GBS also reduces computational effort and decreases the false-positive errors from multiple hypothesis testing [28]. Nevertheless, designing a custom SNP array can be a costly and lengthy process, with genotyping reaching cost-effectiveness at medium to large volumes (thousands of samples). There can also be ascertainment bias introduced when SNPs are selected for the array [103]. In summary, SNP arrays provide high quality, robust SNPs, with ease of data analysis, but potentially at a cost higher than most GBS platforms if sample numbers are low and with some biases.

RRS is the currently most cost-effective genotyping platform for low sample numbers, but suffers from several drawbacks. Polymorphisms in the restriction enzyme recognition site may prevent cutting and lead to erroneous genotyping, so-called allele drop-out. A further issue is the variance in coverage depth between loci, which can be caused by an amplification bias towards shorter fragments with greater GC content. PCR amplification during library preparation can also be uneven, leading to a bias towards specific alleles. These biases do not apply to skim sequencing, and although this platform is costlier than RRS, it is capable of detecting substantially more SNPs, making it more suitable for genome assembly

and validation. The limitations of low coverage skim sequencing are lower rates of SNP genotyping and increased false-positive rates. However, SNP discovery rates and accuracy can be substantially increased using high quality parental genomes, higher sample size, deeper sequencing, filtering, and imputation [14, 104]. Skim sequencing currently remains perhaps the costliest genotyping method, but allows the least biased and most informative sampling of the genome.

5 Conclusion and Perspectives

Genotyping in the era of genomics is now allowing faster, cheaper, more informative, and higher-throughput characterization of crop genomes. While SNP assays provide high quality and the convenience of simple data analysis, the lower cost of RRS and greater data volume produced by skim sequencing suggest that use of GBS will become more prevalent in crop genomics, especially as sequencing costs continue to decrease and more bioinformatics tools are developed to simplify data analysis. New long-read sequencing technologies such as Oxford Nanopore Technologies and Pacific Biosciences are also becoming more cost-competitive and less error-prone [105, 106]. These sequencing platforms may help overcome the challenges associated with short reads, which particularly in complex plant genomes are harder to map. We expect genotyping methods to profit more from long-read sequencing through higher accuracy in the near future. These developments in genotyping platforms and the reducing cost of sequencing will finally help bridge the gap between basic science in plant genomics and applied plant breeding. On the basis of genotyping data, plant breeders will be able to introgress complex agronomic traits into crop germplasm, ensuring robust and nutrient-efficient crops in an age of climate change and increasing food demand [107, 108].

Acknowledgements This work was supported by the University of Western Australia and the Australian Research Council (Projects LP130100925 and LP110100200). The authors thank Cindy Lawley for advice on Illumina array availability.

References

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. *Trends Genet* 30(9):418–426. <https://doi.org/10.1016/j.tig.2014.07.001>
2. Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. *Heredity (Edinb)* 101(1):5–18. <https://doi.org/10.1038/hdy.2008.35>
3. Voss-Fels K, Snowdon RJ (2016) Understanding and utilizing crop genome diversity via high-resolution genotyping. *Plant Biotechnol J* 14(4):1086–1094. <https://doi.org/10.1111/pbi.12456>

4. Edwards D, Forster JW, Chagné D, Batley J (2007) What are SNPs? In: Oraguzie DNC, Rikkerink DEHA, Gardiner DSE, De Silva DHN (eds) Association mapping in plants. Springer, Heidelberg, pp 41–52
5. Ruperao P, Edwards D (2015) Bioinformatics: identification of markers from next-generation sequence data. *Methods Mol Biol* 1245:29–47. https://doi.org/10.1007/978-1-4939-1966-6_3
6. Varshney RK, Singh VK, Hickey JM, Xun X, Marshall DF, Wang J, Edwards D, Ribaut J-M (2015) Analytical and decision support tools for genomics-assisted breeding. *Trends Plant Sci* 21(4):354–363. <https://doi.org/10.1016/j.tplants.2015.10.018>
7. Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol* 12(6):e1001883. <https://doi.org/10.1371/journal.pbio.1001883>
8. Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J* 8(1):2–9. <https://doi.org/10.1111/j.1467-7652.2009.00459.x>
9. Edwards D, Batley J, Snowden RJ (2013) Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 126(1):1–11. <https://doi.org/10.1007/s00122-012-1964-x>
10. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3(10):e3376. <https://doi.org/10.1371/journal.pone.0003376>
11. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res* 17(2):240–248. <https://doi.org/10.1101/gr.5681207>
12. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res* 19(6):1068–1076. <https://doi.org/10.1101/gr.089516.108>
13. Bayer PE, Ruperao P, Mason AS, Stiller J, Chan C-KK, Hayashi S, Long Y, Meng J, Sutton T, Visendi P, Varshney RK, Batley J, Edwards D (2015) High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. *Theor Appl Genet* 128(6):1039–1047. <https://doi.org/10.1007/s00122-015-2488-y>
14. Golicz AA, Bayer PE, Edwards D (2015) Skim-based genotyping by sequencing. *Methods Mol Biol* 1245:257–270. https://doi.org/10.1007/978-1-4939-1966-6_19
15. Kale SM, Jaganathan D, Ruperao P, Chen C, Punna R, Kudapa H, Thudi M, Roorikiwal M, Katta MAVSK, Doddamani D, Garg V, Kishor PBK, Gaur PM, Nguyen HT, Batley J, Edwards D, Sutton T, Varshney RK (2015) Prioritization of candidate genes in “QTL-hotspot” region for drought tolerance in chickpea (*Cicer arietinum* L.) *Sci Rep* 5(5):15296. <https://doi.org/10.1038/srep15296>
16. Rowan BA, Patel V, Weigel D, Schneeberger K (2015) Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3: Genes Genom Genet* 5(3):385–398. <https://doi.org/10.1534/g3.114.016501>
17. Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q (2010) Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc Natl Acad Sci U S A* 107(23):10578–10583. <https://doi.org/10.1073/pnas.1005931107>
18. Barchi L, Lanteri S, Portis E, Valè G, Volante A, Pulcini L, Ciriacci T, Acciarri N, Barbierato V, Toppino L, Rotino GL (2012) A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. *PLoS One* 7(8):e43740. <https://doi.org/10.1371/journal.pone.0043740>
19. Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistue L, Corey A, Filichkina T, Johnson EA, Hayes PM (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics* 12:4. <https://doi.org/10.1186/1471-2164-12-4>

20. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6(5):e19379. <https://doi.org/10.1371/journal.pone.0019379>
21. Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, Lorieux M, Ahmadi N, McCouch S (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor Appl Genet* 126(11):2699–2716. <https://doi.org/10.1007/s00122-013-2166-x>
22. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (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. <https://doi.org/10.1371/journal.pone.0037135>
23. Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7(2):e32253. <https://doi.org/10.1371/journal.pone.0032253>
24. Zhou X, Xia Y, Ren X, Chen Y, Huang L, Huang S, Liao B, Lei Y, Yan L, Jiang H (2014) Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genomics* 15(1):351. <https://doi.org/10.1186/1471-2164-15-351>
25. Wu Z, Wang B, Chen X, Wu J, King GJ, Xiao Y, Liu K (2016) Evaluation of linkage disequilibrium pattern and association study on seed oil content in *Brassica napus* using ddRAD sequencing. *PLoS One* 11(1):e0146383. <https://doi.org/10.1371/journal.pone.0146383>
26. Yu H, Xie W, Wang J, Xing Y, Xu C, Li X, Xiao J, Zhang Q (2011) Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One* 6(3):e17595. <https://doi.org/10.1371/journal.pone.0017595>
27. Zou G, Zhai G, Feng Q, Yan S, Wang A, Zhao Q, Shao J, Zhang Z, Zou J, Han B, Tao Y (2012) Identification of QTLs for eight agronomically important traits using an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods. *J Exp Bot* 63(15):5451–5462. <https://doi.org/10.1093/jxb/ers205>
28. Steemers FJ, Gunderson KL (2007) Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J* 2(1):41–49. <https://doi.org/10.1002/biot.200600213>
29. Bianco L, Cestaro A, Linnsmith G, Muranty H, Denance C, Théron A, Poncet C, Micheletti D, Kerschbamer E, Di Pierro EA (2016) Development and validation of the Axiom® Apple480K SNP genotyping array. *Plant J* 86(1):62–74. <https://doi.org/10.1111/tpj.13145>
30. Vosman B, Pelgrom K, Sharma G, Voorrips R, Broekgaarden C, Pritchard J, May S, Adobor S, Castellanos-Uribe M, van Kaauwen M, Janssen B, van Workum W, Ford-Lloyd B (2015) Phenomics and genomics tools for facilitating brassica crop improvement. *Crop Wild Relat* 10:12–14
31. Roorkiwal M, Jain A, Kale SM, Doddamani D, Chitikineni A, Thudi M, Varshney RK (2017) Development and evaluation of high density SNP Array (Axiom® *Cicer* SNP array) for high resolution genetic mapping and breeding applications in chickpea. *Plant Biotechnol J*. <https://doi.org/10.1111/pbi.12836>
32. Rai KM, Singh SK, Bhardwaj A, Kumar V, Lakhwani D, Srivastava A, Jena SN, Yadav HK, Bag SK, Sawant SV (2013) Large-scale resource development in *Gossypium hirsutum* L. by 454 sequencing of genic-enriched libraries from six diverse genotypes. *Plant Biotechnol J* 11(8):953–963. <https://doi.org/10.1111/pbi.12088>
33. Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA (2012) Development and mapping of SNP assays in allotetraploid cotton. *Theor Appl Genet* 124(7):1201–1214. <https://doi.org/10.1007/s00122-011-1780-8>
34. Pandey MK, Agarwal G, Kale SM, Clevenger J, Nayak SN, Sriswathi M, Chitikineni A, Chavarro C, Chen X, Upadhyaya HD, Vishwakarma MK, Leal-Bertioli S, Liang X, Bertioli DJ, Guo B, Jackson SA, Ozias-Akins P, Varshney RK (2017) Development and evaluation of a high density genotyping ‘Axiom_ *Arachis*’ array with 58K SNPs for accelerating genetics and breeding in groundnut. *Sci Rep* 7:40577. <https://doi.org/10.1038/srep40577>

35. Stoffel K, van Leeuwen H, Kozik A, Caldwell D, Ashrafi H, Cui X, Tan X, Hill T, Reyes-Chin-Wo S, Truco MJ, Michelmore RW, Van Deynze A (2012) Development and application of a 6.5 million feature Affymetrix Genechip[®] for massively parallel discovery of single position polymorphisms in lettuce (*Lactuca* spp.) BMC Genomics 13:185. <https://doi.org/10.1186/1471-2164-13-185>
36. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H, Bertani C, Davassi A, Mayer KFX, Schon CC (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15:823. <https://doi.org/10.1186/1471-2164-15-823>
37. Hill TA, Ashrafi H, Reyes-Chin-Wo S, Yao J, Stoffel K, Truco MJ, Kozik A, Michelmore RW, Van Deynze A (2013) Characterization of *Capsicum annuum* genetic diversity and population structure based on parallel polymorphism discovery with a 30 K unigene pepper GeneChip. PLoS One 8(2):e56200. <https://doi.org/10.1371/journal.pone.0056200>
38. Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun 2:467. <https://doi.org/10.1038/ncomms1467>
39. Yu H, Xie W, Li J, Zhou F, Zhang Q (2014) A whole-genome SNP array (RICE6K) for genomic breeding in rice. Plant Biotechnol J 12(1):28–37. <https://doi.org/10.1111/pbi.12113>
40. Singh N, Jayaswal PK, Panda K, Mandal P, Kumar V, Singh B, Mishra S, Singh Y, Singh R, Rai V, Gupta A, Sharma TR, Singh NK (2015) Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. Sci Rep 5:11600. <https://doi.org/10.1038/srep11600>
41. Koning-Boucoiran CFS, Esselink GD, Vukosavljev M, van 't Westende WPC, Gitonga VW, Krens FA, Voorrips RE, van de Weg WE, Schulz D, Debener T, Maliepaard C, Arens P, Smulders MJM (2015) Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa* L.) Front Plant Sci 6:249. <https://doi.org/10.3389/fpls.2015.00249>
42. Bassil NV, Davis TM, Zhang H, Ficklin S, Mittmann M, Webster T, Mahoney L, Wood D, Alperin ES, Rosyara UR, Koehorst-Vanc Putten H, Monfort A, Sargent DJ, Amaya I, Denoyes B, Bianco L, van Dijk T, Pirani A, Iezzoni A, Main D, Peace C, Yang Y, Whitaker V, Verma S, Bellon L, Brew F, Herrera R, van de Weg E (2015) Development and preliminary evaluation of a 90 K Axiom[®] SNP array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. BMC Genomics 16:155. <https://doi.org/10.1186/s12864-015-1310-1>
43. Lee Y-G, Jeong N, Kim JH, Lee K, Kim KH, Pirani A, Ha B-K, Kang S-T, Park B-S, Moon J-K, Kim N, Jeong S-C (2015) Development, validation and genetic analysis of a large soybean SNP genotyping array. Plant J 81(4):625–636. <https://doi.org/10.1111/tbj.12755>
44. Wang J, Chu SS, Zhang HR, Zhu Y, Cheng H, Yu DY (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6:20728. <https://doi.org/10.1038/srep20728>
45. Winfield MO, Allen AM, BurrIDGE AJ, Barker GLA, Benbow HR, Wilkinson PA, Coghill J, Waterfall C, Davassi A, Scopes G, Pirani A, Webster T, Brew F, Bloor C, King J, West C, Griffiths S, King I, Bentley AR, Edwards KJ (2015) High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. Plant Biotechnol J 14(5):1195–1206. <https://doi.org/10.1111/pbi.12485>
46. Li XH, Han YH, Wei YL, Acharya A, Farmer AD, Ho J, Monteros MJ, Brummer EC (2014) Development of an Alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. PLoS One 9(1):e84329. <https://doi.org/10.1371/journal.pone.0084329>
47. Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M, Salvi S, Jansen J, Viola R, Gut I, Laurens F, Chagne D, Velasco R, van de Weg E, Troglio M (2014) Development and validation of a 20 K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus* × *domestica* Borkh). PLoS One 9(10):e110377. <https://doi.org/10.1371/journal.pone.0110377>

48. Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A, Velasco R, Main D, Rees JD, Iezzoni A, Mockler T, Wilhelm L, Van de Weg E, Gardiner SE, Bassil N, Peace C (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS One* 7(2): e31745. <https://doi.org/10.1371/journal.pone.0031745>
49. Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganai M, Shaw P, Bayer M, Thomas W, Marshall D, Hedley P, Tondelli A, Pecchioni N, Francia E, Korzun V, Walther A, Waugh R (2012) Natural variation in a homolog of *Antirrhinum* CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat Genet* 44 (12):1388–1392. <https://doi.org/10.1038/ng.2447>
50. Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Grainer A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Natl Acad Sci U S A* 103(49):18656–18661. <https://doi.org/10.1073/pnas.0606133103>
51. Close TJ, Bhat PR, Lonardi S, Wu YH, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S, Bozdag S, Roose ML, Moscou MJ, Chao SAM, Varshney RK, Szucs P, Sato K, Hayes PM, Matthews DE, Kleinhofs A, Muehlbauer GJ, DeYoung J, Marshall DF, Madishetty K, Fenton RD, Condamine P, Graner A, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* 10:582. <https://doi.org/10.1186/1471-2164-10-582>
52. Song QJ, Jia GF, Hyten DL, Jenkins J, Hwang EY, Schroeder SG, Osorno JM, Schmutz J, Jackson SA, McClean PE, Cregan PB (2015) SNP assay development for linkage map construction, anchoring whole-genome sequence, and other genetic and genomic applications in common bean. *G3: Genes Genom Genet* 5(11):2285–2290. <https://doi.org/10.1534/g3.115.020594>
53. Snowdon RJ, Luy FLI (2012) Potential to improve oilseed rape and canola breeding in the genomics era. *Plant Breed* 131(3):351–360. <https://doi.org/10.1111/j.1439-0523.2012.01976.x>
54. Dalton-Morgan J, Hayward A, Alameiry S, Tollenaere R, Mason AS, Campbell E, Patel D, Lorenc MT, Yi B, Long Y, Meng JL, Raman R, Raman H, Lawley C, Edwards D, Batley J (2014) A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Funct Integr Genomics* 14(4):643–655. <https://doi.org/10.1007/s10142-014-0391-2>
55. Durstewitz G, Polley A, Plieske J, Luerssen H, Graner EM, Wieseke R, Ganai MW (2010) SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*. *Genome* 53(11):948–956. <https://doi.org/10.1139/G10-079>
56. Delourme R, Falentin C, Fomeju BF, Boillot M, Lassalle G, André I, Duarte J, Gauthier V, Lucante N, Marty A, Pauchon M, Pichon J-P, Ribière N, Trotoux G, Blanchard P, Rivière N, Martinant J-P, Pauquet J (2013) High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. *BMC Genomics* 14(1):120–120. <https://doi.org/10.1186/1471-2164-14-120>
57. Clarke WE, Parkin IA, Gajardo HA, Gerhardt DJ, Higgins E, Sidebottom C, Sharpe AG, Snowdon RJ, Federico ML, Iniguez-Luy FL (2013) Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover sequence nucleotide polymorphisms (SNP) in *Brassica napus* L. *PLoS One* 8(12):e81992. <https://doi.org/10.1371/journal.pone.0081992>
58. Peace C, Bassil N, Main D, Ficklin S, Rosyara UR, Stegmeir T, Sebolt A, Gilmore B, Lawley C, Mockler TC, Bryant DW, Wilhelm L, Iezzoni A (2012) Development and evaluation of a genome-wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. *PLoS One* 7(12):e48305. <https://doi.org/10.1371/journal.pone.0048305>
59. Choudhary S, Gaur R, Gupta S, Bhatia S (2012) EST-derived genic molecular markers: development and utilization for generating an advanced transcript map of chickpea. *Theor Appl Genet* 124(8):1449–1462. <https://doi.org/10.1007/s00122-012-1800-3>

60. Roorkiwal M, Sawargaonkar SL, Chitkineni A, Thudi M, Saxena RK, Upadhyaya HD, Vales MI, Riera-Lizarazu O, Varshney RK (2013) Single nucleotide polymorphism genotyping for breeding and genetics applications in chickpea and pigeonpea using the BeadXpress platform. *Plant Genome* 6(2). <https://doi.org/10.3835/plantgenome2013.05.0017>
61. Gaur R, Jeena G, Shah N, Gupta S, Pradhan S, Tyagi AK, Jain M, Chattopadhyay D, Bhatia S (2015) High density linkage mapping of genomic and transcriptomic SNPs for synteny analysis and anchoring the genome sequence of chickpea. *Sci Rep* 5:13387. <https://doi.org/10.1038/srep13387>
62. Bajaj D, Upadhyaya HD, Khan Y, Das S, Badoni S, Shree T, Kumar V, Tripathi S, Gowda CLL, Singh S, Sharma S, Tyagi AK, Chattopadhyay D, Parida SK (2015) A combinatorial approach of comprehensive QTL-based comparative genome mapping and transcript profiling identified a seed weight-regulating candidate gene in chickpea. *Sci Rep* 5:9264. <https://doi.org/10.1038/srep09264>
63. Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, Saski C, Schnell R, Kuhn D, Motamayor JC (2015) Making a chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*. *DNA Res* 22(4):279–291. <https://doi.org/10.1093/dnares/dsv009>
64. Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, Fang DD, Frelichowski J, Giband M, Hague S, Hinze LL, Kochan KJ, Riggs PK, Scheffler JA, Udall JA, Ulloa M, Wang SS, Zhu QH, Bag SK, Bhardwaj A, Burke JJ, Byers RL, Claverie M, Gore MA, Harker DB, Islam MS, Jenkins JN, Jones DC, Lacape JM, Llewellyn DJ, Percy RG, Pepper AE, Poland JA, Mohan Rai K, Sawant SV, Singh SK, Spriggs A, Taylor JM, Wang F, Yourstone SM, Zheng X, Lawley CT, Ganal MW, Van Deynze A, Wilson IW, Stelly DM (2015) Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. *G3: Genes Genom Genet* 5(6):1187–1209. <https://doi.org/10.1534/g3.115.018416>
65. Close TJ, Lucas MR, Muñoz-Amatriain M, Mirebrahim H, Wanamaker S, Barkley NA, Clair SS, Guo Y-N, Lo S, Huynh BL, Ndeye A, Santos J, Joseph B, Jean-Baptiste T, Drabo I, Kusi F, Atokple I, Boukar O, Fatokun C, Cisse N, Xu P, Roberts P, Lonardi S (2015) A new SNP-genotyping resource for cowpea and its deployment for breeding. In: The plant and animal genome conference, San Diego, January 10–14, 2015
66. Muchero W, Diop NN, Bhat PR, Fenton RD, Wanamaker S, Pottorff M, Hearne S, Cisse N, Fatokun C, Ehlers JD, Roberts PA, Close TJ (2009) A consensus genetic map of cowpea [*Vigna unguiculata* (L.) Walp.] and synteny based on EST-derived SNPs. *Proc Natl Acad Sci U S A* 106(43):18159–18164. <https://doi.org/10.1073/pnas.0905886106>
67. Silva OB, Faria DA, Grattapaglia D (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytol* 206(4):1527–1540. <https://doi.org/10.1111/nph.13322>
68. Myles S, Chia JM, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D (2010) Rapid genomic characterization of the genus *Vitis*. *PLoS One* 5(1):e8219. <https://doi.org/10.1371/journal.pone.0008219>
69. Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schon CC, Sun Q, Walter H, Martin OC, 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. <https://doi.org/10.1371/journal.pone.0028334>
70. Yan JB, Yang XH, Shah T, Sanchez-Villeda H, Li JS, Warburton M, Zhou Y, Crouch JH, Xu YB (2010) High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol Breed* 25(3):441–451. <https://doi.org/10.1007/s11032-009-9343-2>
71. Rousselle Y, Jones E, Charcosset A, Moreau P, Robbins K, Stich B, Knaak C, Flament P, Karaman Z, Martinant JP, Fourneau M, Taillardat A, Romestant M, Tabel C, Bertran J, Ranc N, Lespinasse D, Blanchard P, Kahler A, Chen JL, Kahler J, Dobrin S, Warner T, Ferris R, Smith S (2015) Study on essential derivation in maize: III. Selection and evaluation of a panel of single nucleotide polymorphism loci for use in European and north American germplasm. *Crop Sci* 55(3):1170–1180. <https://doi.org/10.2135/cropsci2014.09.0627>

72. Tian HL, Wang FG, Zhao JR, Yi HM, Wang L, Wang R, Yang Y, Song W (2015) Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties. *Mol Breed* 35(6):136. <https://doi.org/10.1007/s11032-015-0335-0>
73. Tinker NA, Chao SM, Lazo GR, Oliver RE, Huang YF, Poland JA, Jellen EN, Maughan PJ, Kilian A, Jackson EW (2014) A SNP genotyping array for hexaploid oat. *Plant Genome* 7(3). <https://doi.org/10.3835/plantgenome2014.03.0010>
74. Oliver RE, Tinker NA, Lazo GR, Chao SM, Jellen EN, Carson ML, Rines HW, Obert DE, Lutz JD, Shackelford I, Korol AB, Wight CP, Gardner KM, Hattori J, Beattie AD, Bjornstad A, Bonman JM, Jannink JL, Sorrells ME, Brown-Guedira GL, Fetch JWM, Harrison SA, Howarth CJ, Ibrahim A, Kolb FL, McMullen MS, Murphy JP, Ohm HW, Rossnagel BG, Yan WK, Miclauss KJ, Hiller J, Maughan PJ, Hulse RRR, Anderson JM, Islamovic E, Jackson EW (2013) SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species. *PLoS One* 8(3):e58068. <https://doi.org/10.1371/journal.pone.0058068>
75. Tayeh N, Aluome C, Falque M, Jacquin F, Klein A, Chauveau A, Berard A, Houtin H, Rond C, Kreplak J, Boucherot K, Martin C, Baranger A, Pilet-Nayel ML, Warkentin TD, Brunel D, Marget P, Le Paslier MC, Aubert G, Burstin J (2015) Development of two major resources for pea genomics: the GenoPea 13.2K SNP Array and a high-density, high-resolution consensus genetic map. *Plant J* 84(6):1257–1273. <https://doi.org/10.1111/tpj.13070>
76. Deulvot C, Charrel H, Marty A, Jacquin F, Donnadiou C, Lejeune-Henaut I, Burstin J, Aubert G (2010) Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. *BMC Genomics* 11:468. <https://doi.org/10.1186/1471-2164-11-468>
77. Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara UR, Cattonaro F, Vendramin E, Main D, Aramini V, Blas AL, Mockler TC, Bryant DW, Wilhelm L, Troggio M, Sosinski B, Aranzana MJ, Arus P, Jezzone A, Morgante M, Peace C (2012) Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PLoS One* 7(4):e35668. <https://doi.org/10.1371/journal.pone.0035668>
78. Ashrafi H, Hill T, Stoffel K, Kozik A, Yao J, Chin-Wo SR, Van Deynze A (2012) De novo assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for *in silico* discovery of SNPs, SSRs and candidate genes. *BMC Genomics* 13:571. <https://doi.org/10.1186/1471-2164-13-571>
79. Blackmore T, Thomas I, McMahon R, Powell W, Hegarty M (2015) Genetic-geographic correlation revealed across a broad European ecotypic sample of perennial ryegrass (*Lolium perenne*) using array-based SNP genotyping. *Theor Appl Genet* 128(10):1917–1932. <https://doi.org/10.1007/s00122-015-2556-3>
80. Paina C, Byrne SL, Studer B, Rognli OA, Asp T (2016) Using a candidate gene-based genetic linkage map to identify QTL for winter survival in perennial ryegrass. *PLoS One* 11(3):e0152004. <https://doi.org/10.1371/journal.pone.0152004>
81. Studer B, Byrne S, Nielsen RO, Panitz F, Bendixen C, Islam MS, Pfeifer M, Lubberstedt T, Asp T (2012) A transcriptome map of perennial ryegrass (*Lolium perenne* L.) *BMC Genomics* 13:140. <https://doi.org/10.1186/1471-2164-13-140>
82. Plomion C, Bartholome J, Lesur I, Boury C, Rodriguez-Quilon I, Lagravelle H, Ehrenmann F, Bouffier L, Gion JM, Grivet D, de Miguel M, de Maria N, Cervera MT, Bagnoli F, Isik F, Vendramin GG, Gonzalez-Martinez SC (2016) High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). *Mol Ecol Resour* 16(2):574–587. <https://doi.org/10.1111/1755-0998.12464>
83. Gerales A, Difazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N, Porth I, Mckown AD, Skyba O, Li E, Fujita M, Klapste J, Martin J, Schackwitz W, Pennacchio C, Rokhsar D, Friedmann MC, Wasteneys GO, Guy RD, El-Kassaby YA, Mansfield SD, Cronk QCB, Ehrling J, Douglas CJ, Tuskan GA (2013) A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. *Mol Ecol Resour* 13(2):306–323. <https://doi.org/10.1111/1755-0998.12056>

84. Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, Van Deynze A, De Jong WS, Douches DS, Buell CR (2011) Single nucleotide polymorphism discovery in elite north American potato germplasm. *BMC Genomics* 12:302. <https://doi.org/10.1186/1471-2164-12-302>
85. Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CR, Douches DS (2012) Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One* 7(4):e36347. <https://doi.org/10.1371/journal.pone.0036347>
86. Chen H, Xie W, He H, Yu H, Chen W, Li J, Yu R, Yao Y, Zhang W, He Y, Tang X, Zhou F, Deng XW, Zhang Q (2014) A high-density SNP genotyping array for rice biology and molecular breeding. *Mol Plant* 7(3):541–553. <https://doi.org/10.1093/mp/sst135>
87. Zhao KY, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung CW, Reynolds A, Bustamante CD, McCouch SR (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS One* 5(5):e10780. <https://doi.org/10.1371/journal.pone.0010780>
88. Travis AJ, Norton GJ, Datta S, Sarma R, Dasgupta T, Savio FL, Macaulay M, Hedley PE, McNally KL, Sumon MH, Islam MR, Price AH (2015) Assessing the genetic diversity of rice originating from Bangladesh, Assam and West Bengal. *Rice* 8:35. <https://doi.org/10.1186/s12284-015-0068-z>
89. Ye CR, Tenorio FA, Argayoso MA, Laza MA, Koh HJ, Redona ED, Jagadish KSV, Gregorio GB (2015) Identifying and confirming quantitative trait loci associated with heat tolerance at flowering stage in different rice populations. *BMC Genet* 16:41. <https://doi.org/10.1186/s12863-015-0199-7>
90. Thomson MJ (2014) High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed Biotechnol* 2(3):195–212. <https://doi.org/10.9787/PBB.2014.2.3.195>
91. Haseneyer G, Schmutzer T, Seidel M, Zhou RN, Mascher M, Schon CC, Taudien S, Scholz U, Stein N, Mayer KFX, Bauer E (2011) From RNA-seq to large-scale genotyping: genomics resources for rye (*Secale cereale* L.) *BMC Plant Biol* 11:131. <https://doi.org/10.1186/1471-2229-11-131>
92. Bekele WA, Wieckhorst S, Friedt W, Snowdon RJ (2013) High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. *Plant Biotechnol J* 11(9):1112–1125. <https://doi.org/10.1111/pbi.12106>
93. Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One* 8(1):e54985. <https://doi.org/10.1371/journal.pone.0054985>
94. Hyten DL, Song Q, Choi IY, Yoon MS, Specht JE, Matukumalli LK, Nelson RL, Shoemaker RC, Young ND, Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theor Appl Genet* 116(7):945–952. <https://doi.org/10.1007/s00122-008-0726-2>
95. Akond M, Liu S, Schoener L, Anderson JA, Kantartzi SK, Meksem K, Song Q, Wang D, Wen Z, Lightfoot DA, Kassem MA (2013) SNP-based genetic linkage map of soybean using the SoySNP6K Illumina Infinium BeadChip genotyping array. *J Plant Genome Sci* 1(3):80–89. <https://doi.org/10.5147/jpgs.2013.0090>
96. Bachlava E, Taylor CA, Tang SX, Bowers JE, Mandel JR, Burke JM, Knapp SJ (2012) SNP discovery and development of a high-density genotyping array for sunflower. *PLoS One* 7(1):e29814. <https://doi.org/10.1371/journal.pone.0029814>
97. Talukder ZI, Gong L, Hulke BS, Pegadaraju V, Song QJ, Schultz Q, Qi LL (2014) A high-density SNP map of sunflower derived from RAD-sequencing facilitating fine-mapping of the rust resistance gene R-12. *PLoS One* 9(7):e98628. <https://doi.org/10.1371/journal.pone.0098628>
98. Sim SC, Durstewitz G, Plieske J, Wieseke R, Ganai MW, Van Deynze A, Hamilton JP, Buell CR, Causse M, Wijeratne S, Francis DM (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One* 7(7):e40563. <https://doi.org/10.1371/journal.pone.0040563>

99. Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieseke R, Plieske J, International Wheat Genome Sequencing Consortium, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, Morell M, Dubcovsky J, Ganal M, Tuberosa R, Lawley C, Mikoulitch I, Cavanagh C, Edwards KJ, Hayden M, Akhunov E (2014) Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol J* 12(6):787–796. <https://doi.org/10.1111/pbi.12183>
100. Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor Appl Genet* 119(3):507–517. <https://doi.org/10.1007/s00122-009-1059-5>
101. Cavanagh CR, Chao SM, Wang SC, Huang BE, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira GL, Akhunova A, See D, Bai GH, Pumphrey M, Tomar L, Wong DB, Kong S, Reynolds M, da Silva ML, Bockelman H, Talbert L, Anderson JA, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell PL, Dubcovsky J, Morell MK, Sorrells ME, Hayden MJ, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc Natl Acad Sci U S A* 110(20):8057–8062. <https://doi.org/10.1073/pnas.1217133110>
102. Mason AS, Zhang J, Tollenaere R, Vasquez Teuber P, Dalton-Morgan J, Hu L, Yan G, Edwards D, Redden R, Batley J (2015) High-throughput genotyping for species identification and diversity assessment in germplasm collections. *Mol Ecol Resour* 15(5):1091–1101. <https://doi.org/10.1111/1755-0998.12379>
103. Lachance J, Tishkoff SA (2013) SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *BioEssays* 35(9):780–786. <https://doi.org/10.1002/bies.201300014>
104. Fu LX, Cai CC, Cui YN, Wu J, Liang JL, Cheng F, Wang XW (2016) Pooled mapping: an efficient method of calling variations for population samples with low-depth resequencing data. *Mol Breed* 36(4):48–48. <https://doi.org/10.1007/s11032-016-0476-9>
105. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* 33(6):623–630. <https://doi.org/10.1038/nbt.3238>
106. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 12(4):351–356. <https://doi.org/10.1038/nmeth.3290>
107. Abberton M, Batley J, Bentley A, Bryant J, Cai H, Cockram J, Costa de Oliveira A, Cseke LJ, Dempewolf H, De Pace C, Edwards D, Gepts P, Greenland A, Hall AE, Henry R, Hori K, Howe GT, Hughes S, Humphreys M, Lightfoot D, Marshall A, Mayes S, Nguyen HT, Ogonnaya FC, Ortiz R, Paterson AH, Tuberosa R, Valliyodan B, Varshney RK, Yano M (2015) Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnol J* 14(4):1095–1098. <https://doi.org/10.1111/pbi.12467>
108. Batley J, Edwards D (2016) The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Curr Opin Plant Biol* 30:78–81. <https://doi.org/10.1016/j.pbi.2016.02.002>

Trait Mapping Approaches Through Linkage Mapping in Plants



Pawan L. Kulwal

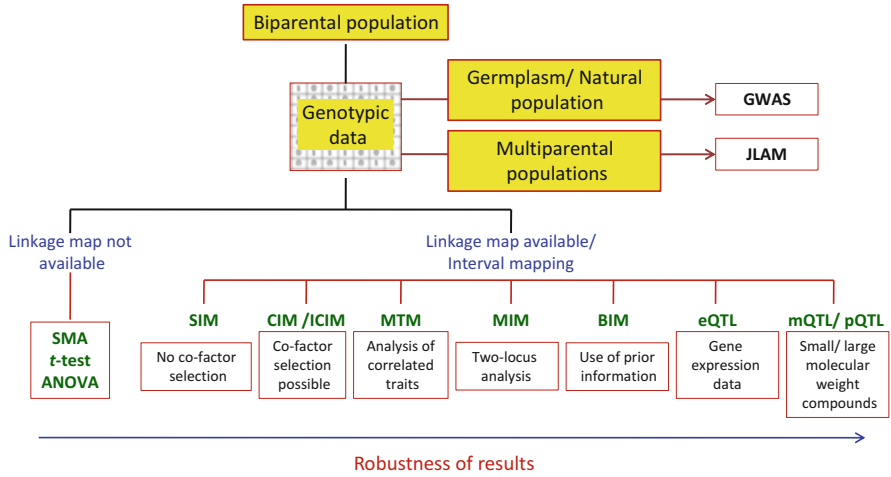
Abstract Quantitative trait loci (QTL) mapping in crop plants has now become a common practice due to the advances made in the area of molecular markers as well as that of statistical genomics. Consequently, large numbers of QTLs have been identified in different crops for a variety of traits. Several computational tools are now available that suit the type of mapping population and the trait(s) to be studied for QTL analyses as well as the objective of the program. These methods are comprised of simpler approaches like single marker analysis and simple interval mapping to relatively exhaustive inclusive composite interval mapping and Bayesian interval mapping. The relative significance of each of these methods varies considerably. The progress made in the area of computational analysis involving the identification of QTLs either through interval mapping or association mapping is unprecedented, and it is expected that it will continue to evolve over the coming years. An overview of the different methods of linkage-based QTL analysis is provided in this chapter.

P. L. Kulwal (✉)

State Level Biotechnology Centre, Mahatma Phule Agricultural University, Rahuri,
Maharashtra, India

e-mail: pawankulwal@gmail.com

Graphical Abstract



Keywords Interval mapping, Marker-trait association, Plants, QTL mapping

Contents

1	Introduction	56
2	Methods of Linkage-Based QTL Mapping	56
2.1	Identification of Marker-Trait Association (MTA) When Linkage Map Is Not Available: Single-Marker Analysis	57
2.2	Identification of QTL When Linkage Map Is Available	58
2.3	Identification of Interacting or Epistatic QTLs: Two-Locus Analysis	59
2.4	Mapping QTL for Correlated Traits Simultaneously	61
2.5	Mapping QTL Using Prior Information: A Bayesian Approach for QTL Mapping ...	62
2.6	The Analysis of Traits for Which Data Are Recorded Periodically: QTL Mapping for Dynamic Traits	63
2.7	Analysis of Traits for Which Data Are Scored on a Numeric Scale: QTL Mapping for Ordinal Traits	64
2.8	Meta-QTL Analysis	65
2.9	Mapping of QTLs for Gene Expression and for Large and Small Molecular Weight Compounds: The Concept of Genetical Genomics	65
2.10	Identification of QTLs Using Multiparental Mapping Populations: Joint Linkage-Association Mapping	67
2.11	Quantitative Resistance Loci (QRLs) Governing Quantitative Disease Resistance (QDR)	68
2.12	Discovery and Introgression of Useful QTLs from Wild-Type or Unadapted Germplasm: Advanced Backcross QTL Analysis	68
3	Factors Affecting Results of QTL Mapping in Plants	69
3.1	Heritability of the Trait	69
3.2	Size and Nature of Mapping Population	70
3.3	Number of Markers in the Linkage Map	71

3.4 Method of Analysis	71
4 Computer Programs for QTL Analysis	72
5 Conclusion and Outlook	72
References	74

Abbreviations

AB-QTL	Advanced backcross QTL
AM	Association mapping
BSA	Bulk segregant analysis
CIM	Composite interval mapping
DH	Doubled haploid
EM algorithm	Expectation maximization algorithm
eQTL	Expression QTL
GLM	General linear model
GS	Genomic selection
GWAS	Genome wide association studies
ICIM	Inclusive composite interval mapping
IM	Interval mapping
JLAM	Joint linkage-association mapping
LD	Linkage disequilibrium
MAGIC	Multi-parent advanced generation intercross
MAS	Marker-assisted selection
MCILs	Multiline cross-inbred lines
MIM	Multiple interval mapping
M-QTL	Main effect QTL
mQTL	Metabolite QTL
MTA	Marker-trait association
MTMIM	Multiple-trait multiple interval mapping
NAM	Nested association mapping
pQTL	Protein QTL
QDR	Quantitative disease resistance
QE	QTL \times environment interactions
QQ	QTL \times QTL interactions
QQE	QTL \times QTL \times environment interactions
QRL	Quantitative resistance loci
QTL	Quantitative trait loci
RIAILs	Recombinant inbred advanced intercross lines
SIM	Simple interval mapping
SMA	Single marker analysis
TFM	Time-fixed mapping
TRM	Time-related mapping

1 Introduction

Understanding the genetics of quantitative traits has been a common focus over the last few decades. Ever since Sax [1] demonstrated the use of a simple t -test for finding the association between the seed weight and color of beans, methods of mapping quantitative trait loci (QTL) in plants have evolved steadily over the years. However, over last three decades there has been a renewed interest in studying the genetics of these traits due to the availability of large numbers of genomic resources including mapping populations, molecular markers, linkage maps, and computational tools. The progress in this area has been quite unprecedented. Consequently, large numbers of statistical methods are now available that suit the nature of the trait and mapping population as well as the objective of the research. As a result, it has now become possible to rapidly identify QTLs as well as candidate genes associated with individual traits. A large number of marker-trait associations (MTAs) for different traits have also been identified in different crops, and several of these have been deployed successfully in crop improvement programs through marker-assisted selection (MAS) [2, 3]. Some of the QTLs identified over the years have also been cloned successfully in different crop plants [2, 4, 5]. Similarly, the literature regarding this aspect has also grown tremendously. Many of reviews describing different methods of QTL analysis and its various dimensions, with special emphasis on crop plants, have appeared over the years [2, 6–15]. A partial list of references on statistical genetics is available at <http://pages.stat.wisc.edu/~yandell/statgen/reference/software.html>.

In this chapter, the different methods of QTL analysis that are based on the principle of linkage are discussed without describing much of the statistics involved in it (Fig. 1). Comparison between these different methods, factors affecting them, and the recent trends in the QTL analysis in crop plants are also discussed, along with different computer programs available for analysis of the data. However, the aspect of association mapping, which is based on the principle of linkage disequilibrium (LD), is not covered here, but is available in another chapter in this book.

2 Methods of Linkage-Based QTL Mapping

The different methods of linkage-based QTL analysis can be divided into four main categories depending upon the principle involved in it, and can be classified as (1) single-marker analysis when linkage map is not available, (2) interval mapping when linkage map is available, (3) meta-QTL analysis and (4) joint linkage and association mapping. Accordingly, these different methods are discussed in the following sections.

In plant-breeding experiments, data on a trait are recorded in various ways (during growth stages, at maturity) either in a continuous scale or in several ordered categories. Therefore based on the nature of trait being studied, the different methods of QTL mapping have also been discussed in this section.

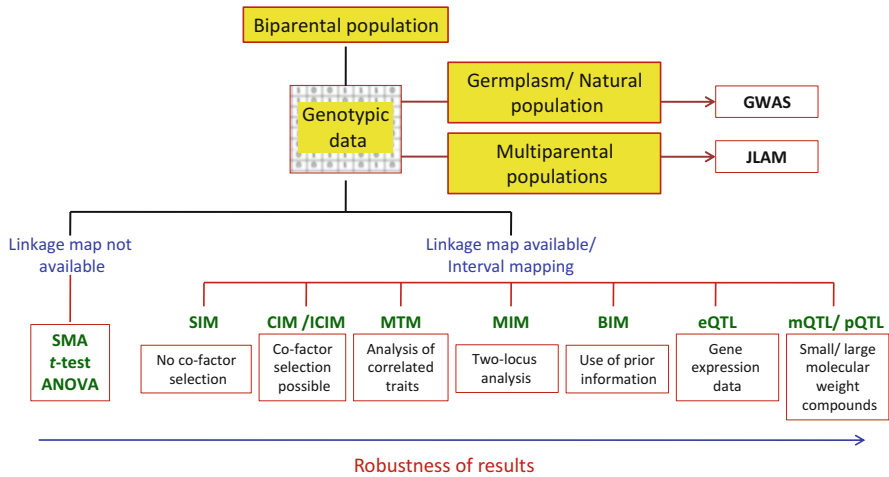


Fig. 1 Pictorial representation of different methods of QTL analysis. It describes that when a biparental population is genotyped with molecular markers and genotypic data is available but linkage map is not available, one can use single marker analysis (SMA) (t-test or ANOVA) for identification of marker-trait associations. When linkage map is available, one can analyze the data through simple interval mapping (SIM), composite/inclusive composite interval mapping (CIM/ICIM), multiple trait mapping (MTM), multiple interval mapping (MIM), Bayesian interval mapping (BIM), expression QTL (eQTL) or metabolite or protein QTL (mQTL/pQTL). The criteria used in each of these interval mapping approaches are given in the box below the method. The relative robustness of results of these methods over one another is shown with arrow. When germplasm/natural population is genotyped, one can perform genome wide association study (GWAS), while multiparental populations enable joint linkage and association mapping (JLAM)

2.1 Identification of Marker-Trait Association (MTA) When Linkage Map Is Not Available: Single-Marker Analysis

During the initial years when limited numbers of marker resources were available and statistical programs for development of linkage maps and interval mapping were still in their infancy, MTAs were identified based on rather simple approaches. The approach of bulk segregant analysis (BSA) proposed by Michelmore et al. [16] was very commonly used. In this approach, molecular marker(s) showing polymorphism between the parental genotypes of the mapping population and the two pools or bulks of DNA samples differing in a trait of interest are first selected and subjected to BSA for further selection of putative markers, which are then used for genotyping the whole mapping population. The putative QTL can thus be detected from the analysis of such markers following any of the single-marker analysis (SMA) methods. It is still considered as a rapid approach (short cut) for detecting the linkage of a marker with a QTL for a trait of interest. Several important QTLs that were earlier identified using BSA were later confirmed following advance methods of interval mapping. The advantage with this method is that the huge cost often incurred in genotyping the entire population could be saved.

Although proposed more than two decades ago, the approach still remains popular among the scientific community for quick analysis of the data. Large numbers of studies have used the principle of BSA and identified important QTLs for various traits in different crop plants. Recently, BSA was used for the identification of major grain yield QTLs under drought stress in rice [17]. Similarly, in another study using the whole genome-resequencing approach (also called QTL-seq) in rice, two bulks comprising 20–50 individuals with extreme phenotypic values were analyzed and QTLs for important agronomic traits were identified [18]. Although initially proposed to be used in biparental populations, the principle of DNA pooling from extreme genotypes for the rapid identification of QTLs has also seen application in an association mapping experiment. Using this approach, recently Kujur et al. [19] identified three major QTLs and candidate genes for seed weight in chickpea. Because of its simple, time- and cost-effective features, BSA still holds promise in the QTL-mapping programs.

Different methods commonly used for SMA include the *t*-test, ANOVA, and simple regression [7, 20].

(i) *t*-test, ANOVA, or regression approach: One of the simplest ways to determine whether an association exists between a molecular marker and the trait of interest is to conduct a single-factor analysis of variance (ANOVA). In this method the marker and the trait of interest are considered as independent and dependent variables, respectively. The marker-trait association (MTA) is considered only if the marker under consideration shows a significant difference between the two marker classes for the trait of interest. Based on this simple analysis, a QTL can be inferred to be located adjoining to, or in the vicinity of, the identified marker. Similarly, linear regression can be used for the identification of MTA and can help in estimating the phenotypic variation arising from the QTL linked to the marker. The advantage with this approach is that it is computationally very easy and can be performed even when one does not have a linkage map available. Often such types of situations arise when sufficient markers are not available, which limits the construction of a linkage map. However, the major drawback with this method is that the further a QTL is from a marker, the less likely it will be detected. Several QTL mapping studies in crop plants have utilized this approach for the identification of QTLs for a variety of traits. Many of these QTLs were subsequently confirmed using the approach of interval mapping.

2.2 Identification of QTL When Linkage Map Is Available

The era of development of framework linkage maps and interval mapping in plants began with the availability of interactive computer package MAPMAKER [21, 22]. Ever since its availability, it has been by far the most used computer program for the development of linkage maps. It not only provided the basis for framework maps, but it also introduced the principle of simple interval mapping (SIM) for the mapping of QTL by scanning an interval between each pair of

markers in the genome. Not only did it facilitate QTL identification, but it also addressed the shortcomings of SMA. During the 1990s, majority of the QTL mapping studies were carried out using the principle of SIM. It was only when the principle of combining IM with multiple regression was introduced [23–25] that the problems of SIM were addressed. This method was later named “composite interval mapping” (CIM; [25]). This was a significant development and changed the way QTL mapping studies used to be carried out. CIM became the method of choice and by far the most popular QTL mapping approach amongst the scientific community. In order to avoid chances of false-positive associations and to increase the efficiency of QTL detection, improvements in the form of empirical threshold and permutation tests have also been proposed [26, 27]. Another method called inclusive composite interval mapping (ICIM), which fixed the problem of arbitrary cofactor selection in CIM, was later proposed by Li et al. [28]. The advantage with this method is that it takes into account the significant cofactors and calculates their effects using stepwise regression before IM is conducted and the effects are fixed during genome scanning. This method has been found to improve QTL detection efficiency over that of CIM and has been used in many studies. Interval mapping can be accomplished using any of the available methods including SIM, CIM, ICIM, and several variants proposed later. Comparison between different methods of QTL analysis is given in Table 1.

Several variants of QTL mapping were proposed subsequent to CIM that offered a better understanding of the genetics of complex traits. These include studies of multiple marker intervals simultaneously and identification of epistatic (interacting) QTLs (multiple interval mapping, MIM), analysis of multiple traits simultaneously taking into account trait correlations, analysis of dynamic and ordinal traits, and many more. These methods are discussed in greater detail in the following sections. Some of these methods, despite once being considered computationally intensive, are being used on a regular basis due to advances in computational tools. Large numbers of QTL mapping studies using either of these approaches have been conducted in different crop plants and it is not possible to include all of them in this chapter.

2.3 Identification of Interacting or Epistatic QTLs: *Two-Locus Analysis*

The principle of epistasis has been known to geneticists for a long time and its importance in plant breeding has been well documented [29]. However, only QTLs having a main effect (M-QTL) were used for identification in the majority of the earlier studies (single-locus analysis). This was mainly because of the computational complexity involved with using multiple QTLs in the statistical model [30]. This becomes more complex if higher order interactions are involved [31]. It therefore did not allow the identification of interacting QTLs (QTL \times QTL;

Table 1 Comparison between different methods of QTL analysis

Particular	SMA	SIM	CIM	ICIM	MIM	MTM	BIM	AM	JLAM
Linkage map	Not required	Required	Required	Required	Required	Required	Required	Not required	Required
Population requirement	Biparental	Biparental	Biparental	Biparental	Biparental	Biparental	Biparental	Germplasm; breeding lines	Multiparental population
Cofactors taken into account	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Exact position of QTL can be find out	No	No (not precise)	Yes	Yes	Yes	Yes	Yes	No/Yes (depends on availability of map)	Yes
Computationally intensive	Least	Least	Moderate	Moderate	High	High	High	High	High
False positive	More	More	Less	Less	Less	Less	Least	Low/high (depend on significance criteria used)	Less
Analysis of correlated traits	Not possible	Not possible	Not possible	Not possible	Not possible	Possible	Not possible	Not possible	Not possible
Marker information utilized during analysis	Single marker	Flanking marker	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome
Precision of QTL effect	No	Not precise as that of CIM	Precise	Precise	Precise	Precise	Precise	Precise	Precise
Ability to incorporate prior information	Not possible	Not possible	Not possible	Not possible	Possible	Possible	Possible	Possible ^a	Possible ^a

^aPossible only if the method involves use of Bayesian statistics

two-locus analysis). It is also logical to think that there may be QTLs that may or may not have a main effect, but can interact with another such QTL [32, 33]. These types of interacting QTLs also contribute significantly to trait variation. Therefore, it was also thought appropriate to put the principle of epistasis into QTL interval mapping. Accordingly, multiple interval mapping (MIM) was proposed by Kao et al. [34]. Similarly, in another study, a mixed model approach was proposed by Wang et al. [30] that enabled the identification of not only QTL \times QTL (QQ) interactions, but also QTL \times environment (QE), and QTL \times QTL \times environment (QQE) interactions. Very recently, a three-stage search strategy for the mapping of epistatic QTLs has been proposed by Laurie et al. [35]. In this approach, first the main effect QTLs are identified, which is followed by the identification of epistatic QTLs interacting significantly with other QTLs, and, finally, new epistatic QTLs are searched in pairs. These methods not only improved the precision of the commonly used approach of CIM, but also increased the efficiency of QTL-mapping experiments, as interacting QTLs (QQ and QE) that contribute significantly to the total variation of the trait could be identified. These approaches have also been included in the commonly used QTL-mapping software: QTL Network and QTL Cartographer [30, 36]. A large number of studies involving identification of such interactions have now been carried out in different crops including rice [30, 37, 38], wheat [33, 39–41], maize [42, 43], and barley [44, 45]. It was also shown that in wheat the proportion of variation explained by QQ and QE or QQE varies from trait to trait [39].

Molecular marker-based QTL mapping studies have provided more evidence for epistasis than the conventional biometric approaches of quantitative genetics. Therefore, for long-term progress in plant breeding, one cannot ignore the importance of epistasis [29]. In order to completely dissect the trait in terms of its total variation, it is imperative that these interactions, including higher order interactions, be identified [31]. However, the methodology for addressing the issue of higher order interactions is still underdeveloped.

2.4 Mapping QTL for Correlated Traits Simultaneously

It is a common practice to conduct QTL analysis separately for each trait. However, it is often observed that some of the traits are significantly correlated with each other. The ability to identify and use a common QTL governing more than one trait can accelerate and increase the efficiency of MAS programs significantly. Multiple-trait QTL analysis is QTL analysis applied to several traits simultaneously and can help in the identification of pleiotropic QTLs. The importance of such pleiotropic QTLs and multi-trait QTL analysis has earlier been advocated and also empirically demonstrated [46–48]. Taking into account the correlation structure among the traits, this type of analysis was shown to improve the statistical power of QTL detection and the precision of parameter estimation in these studies. Later this approach was also incorporated into the popularly used QTL analyses program

“QTL Cartographer” and other software, and was also successfully used in wheat [40, 49], sorghum [50], and other crops for different traits.

Recently an improvement over the existing method of multiple-trait analysis was proposed by Silva et al. [51], which takes into account the genetic and environmental correlations between traits and provides more details on the genetic architecture of complex traits by separating pleiotropic QTLs from closely linked non-pleiotropic QTL and QE interactions. Further, it can also estimate the total genotypic variance-covariance matrix between the correlated traits and decompose it in terms of QTL-specific variance-covariance matrices. It is expected that this method of multiple-trait multiple-interval mapping (MTMIM) of correlated traits will be more rewarding and can enhance the speed of MAS.

2.5 Mapping QTL Using Prior Information: A Bayesian Approach for QTL Mapping

In genetics, Bayesian analysis has been used for a long time and has now become an integral part of the QTL mapping studies. It is always said that statistics deals with uncertainty that is relative to the information we have [52]. In other words, the less information, the more uncertainty, and vice versa. As opposed to the commonly used methods of QTL analysis (SMA, SIM, and CIM), also called frequentist methods, which depend on the fixed parameters, Bayesian analysis deals with the uncertainty of the data based on prior information that is gathered and updated regularly to draw the posterior distribution according to Bayes’ rule. It therefore allows for easy and systematic incorporation of prior knowledge into the data analysis [53]. Accordingly, a Bayesian model consists of three components: (1) prior distribution, (2) conditional distribution, and (3) posterior distribution.

Although once considered to be computationally demanding, in recent years the Bayesian application has become an integral part of not only QTL analysis experiments, but also of association mapping [54, 55] and genomic selection (GS) experiments [56, 57]. This all has been made possible due to advances in the computational methodologies over the last few years. In one of its earliest demonstrated uses in QTL mapping, Satagopan et al. [58] used the Bayesian principle for estimating the locations and effect parameters for multiple QTLs with pre-specified numbers of QTLs in a DH progeny of *Brassica napus*. Since then, a large number of studies on crop plants involving the principles of Bayesian statistics have been published and it has now become an almost integral part of any analyses.

With the growing interest in this approach, new models were also proposed that facilitated the analysis of binary and ordinal traits [59, 60], interacting QTLs/epistasis [61–63], permutation testing [64], QE interactions [65], multiple QTL analysis [66], multiple trait analysis [67, 68], and pleiotropy [69]. The complexity of identifying epistatic QTLs, appropriate model selection, and many other issues

that earlier plagued the efficient analysis of QTLs were addressed in these studies. The only concern that might limit the use of the Bayesian approach in analysis is that different conclusions can be drawn by different researchers if they use different priors in their analysis [2, 70]. Notwithstanding this, Bayesian statistics is the preferred choice of the statistician and will be used for a long time in all aspects of genetic analysis.

2.6 The Analysis of Traits for Which Data Are Recorded Periodically: QTL Mapping for Dynamic Traits

In majority of the QTL mapping studies, the data on a quantitative trait measured at a fixed time point or stage of growth/ontogenesis are used for analysis. This way of analyzing the data can identify QTLs and estimate their effects, which are accumulated over time from the beginning of growth until the time of actual observation. However, it is a well-known fact that the development of a trait is an end result of differential activities of many related QTLs, which express during the life cycle of the crop. This is because the developmental traits are under the control of genes, which are expressed at specific stages of development in response to the existing environmental conditions. Therefore, the traits for which phenotypic values change over time during the period of growth are called dynamic traits. Wu et al. [71] called the QTL mapping of such traits time-related mapping (TRM), as opposed to time-fixed mapping (TFM) for the traits for which the data are recorded at a fixed time or stage. Later, Wu and Lin [72] termed this aspect “functional mapping.” The advantage with this approach is that recorded observations of the same individuals over different developmental stages are a form of replication that can increase the statistical power of QTL detection. Besides this, another important advantage of this approach is that the stage of growth at which the heritability of the trait is highest can also be identified. The QTLs identified at this stage will be more useful for a breeding program involving MAS [73]. One of the very common examples of this is plant height in crops, for which the differences are visible during early growth but are neutralized/minimized towards maturity.

Several QTL mapping studies have been carried out for dynamic traits in different crop plants and have reported some common as well as growth-stage specific QTLs. Earlier this approach was successfully used in rice to identify QTLs associated with increased grain filling percentage per panicle [74]. Similarly, dynamic QTLs for seed reserve utilization were identified during three germination stages in rice [75]. It was observed that more QTLs express at the late germination stage. Osman et al. [76] used this approach along with conditional analysis for growth and yield traits under submergence conditions in maize and identified some common and some stage-specific QTLs. Similarly, in a recent study in triticale a population comprised of 647 doubled haploid lines derived from four families were phenotyped for plant height using a precision phenotyping platform at multiple time

points. The study identified main effect and epistatic QTLs for plant height for each of the time points. Some of these QTLs were detected at all time points whereas others were specific to particular developmental stages, while the contribution of the QTL to the genotypic variance of plant height also varied with time [77]. Recently, a Bayesian nonparametric approach was also proposed for the analysis of dynamic traits [78], which offers advantages over the existing methods of analysis. The only limitation of this method is that it cannot be used for traits on which periodical observations are not possible (for example, grain protein content, grain yield, etc.).

2.7 Analysis of Traits for Which Data Are Scored on a Numeric Scale: QTL Mapping for Ordinal Traits

QTL analysis of the trait is based on the data that are recorded on a continuous scale with the assumption that they show normality. However, in nature, many quantitative traits in plants like disease resistance or quality parameters are recorded on a certain scale in several ordered categories based on intensity or severity. Although these traits are quantitative in nature, the data do not show continuous variation and therefore contain less information. These types of traits are called ordinal traits, and appropriate statistical treatment is required to deal with this type of trait distribution. Nevertheless, in many earlier published reports of QTL mapping, data on ordinal traits was analyzed in the same way as that of continuous traits. One of the reasons attributed for treating these traits similarly in earlier studies was partly the lack of availability of statistical tools to deal with these traits. However, QTL mapping methods for dealing with ordinal traits have evolved over the years, with more emphasis on traits studied in humans than in plants.

Earlier methods for QTL analysis of ordinal traits in back-crossed populations using the general linear model (GLM) were proposed by Hackett and Weller [79], and Xu and Atchley [80], which was later extended to four-way crosses by Rao and Xu [81]. An improvement over the existing GLM method was later proposed by Xu and Xu [82] in the form of a multivariate model to deal with the ordinal traits based on the EM algorithm. Subsequently, the principle of MIM described earlier for continuous traits was also extended to ordinal and binary traits for the identification of multiple QTL effects and epistasis [83]. This method is also included in the popular QTL analysis program QTL Cartographer. More recently, another approach based on an efficient hierarchical GLM was proposed for the identification of main-effect QTL and QE interactions governing ordinal traits in AM experiments [84].

2.8 *Meta-QTL Analysis*

During the last two decades, there has been a surge in the number of QTL mapping studies in different crop plants, which has resulted in several thousand published articles (source, Google Scholar). It is also seen that QTL mapping for the same traits are carried out in different genetic backgrounds in the same crop, leading to the identification of several QTLs. It thus necessitates the integration of QTL mapping results from these individual experiments performed on the same crop to identify common as well as novel loci/alleles underlying complex traits, for their effective use in crop improvement programs [2]. Meta-analysis of QTLs is an important approach that integrates information from multiple QTL-mapping studies and allows greater statistical power for QTL detection and more precise estimation of their genetic effects. Besides this, meta-QTL analysis can help to refine the genomic regions of interest frequently identified in different studies, and can provide the closest flanking markers [85]. Hence, a meta-analysis can be more rewarding than those of individual studies and can give greater insight into the genetic architecture of complex traits [86].

Because of its ability to integrate results from several individual QTL mapping studies, this approach has been used in many crops, and several meta-QTLs have also been identified. In one of the first examples, Chardon et al. [87] used the approach of Goffinet and Gerber [88] to study the genetic basis of flowering time in maize by integrating results of several mapping studies. From the total of 313 QTLs used for the study, they identified a total of 62 consensus QTLs and also reported a twofold increase in the precision of QTL position estimation from the original one. Several such studies were later carried out in different crops, including: disease resistance in cocoa [89]; fiber quality, yield, and biotic and abiotic stress tolerance in cotton [90, 91]; drought tolerance in rice [92]; late blight resistance and plant maturity traits in potato [85]; root genetic architecture in rice [93] and maize [94]; and protein concentration in soybean [95]. A list of several such studies carried out in cereals is also given in Gupta et al. [2]. These studies have also been made computationally possible due to the availability of software tools like BioMercator [96] and MetaQTL [97].

2.9 *Mapping of QTLs for Gene Expression and for Large and Small Molecular Weight Compounds: The Concept of Genetical Genomics*

As is the case with many physiological traits, variation in gene expression (m-RNA), as well as that of large and small molecular weight compounds (protein or metabolite), often shows a quantitative distribution, thereby allowing its genetic dissection using the commonly used methods of QTL mapping [98]. Earlier, the term genetical genomics was restricted only to the mapping of expression QTL

(eQTL) [99]. However, the last decade has seen tremendous progress in terms of cost-effective high-throughput genotyping techniques, which made it possible to study the complexity of traits by measuring not only gene expression, but also thousands of proteins and metabolites to map eQTL, protein QTL (pQTL), and metabolite QTL (mQTL), respectively [100–102]. In the experiments involving genetical genomics, data on gene expression or individual proteins or metabolites can be used as a phenotype in QTL analysis. The large-scale data on gene expression (genetical genomics), if combined with genetics, can help in connecting phenotypic variation to genotypic diversity and can lead to the identification of genetic regulatory loci, and ideally genes, which explain the observed variation [98]. The rationale behind this approach is that a specific gene's expression level is easier to quantify than the more complex developmental or physiological traits. Thus, if the loci governing differential gene expression patterns is identified and compared with that of the loci controlling a specific physiological trait, one can have better understanding of the complex traits [103]. It is thus obvious that integration of omics data in genetic studies can reduce the number of candidate genes for a given QTL from hundreds to a sizeable list [98].

The earlier studies on genetical genomics predominantly utilized microarrays for the analysis of mapping populations in a variety of species. However, experiments involving microarrays are very expensive, thereby limiting their use in all such studies. Metabolomics platforms on the other hand are much cheaper per sample than transcriptomics, enabling large populations to be studied with sufficient replication for moderate-to-low heritability traits. Moreover, most metabolomics platforms are higher-throughput than transcriptomics, allowing for rapid analysis. Therefore, in recent years there are increasing numbers of reports pertaining to mQTL analysis in plants. Some of them have been described elsewhere ([2, 104]; also see Alseekh et al. [105]). Although earlier these studies were more common in model species like *Arabidopsis* ([106] and references therein [107]), they are also being carried out in different crops including potato [108], brassica [109], tomato, and wheat. Very recently, a comprehensive mQTL analysis was carried out in tomato, and a total of 679 mQTLs for secondary metabolism in tomato fruit pericarp were detected in 76 introgression lines [105]. Similarly, in wheat, mQTL analysis was combined with that of QTL analysis for agronomic traits in a doubled haploid population [110]. These studies are not limited to biparental populations, but are also becoming very popular in AM experiments (for details, see Luo [111]).

Genetical genomics has offered lots of understanding about the influence of genetic factors on a biological system. However, as like any quantitative trait, molecular networks are also influenced by environmental conditions. Therefore, for a better and complete understanding of these networks, it is necessary that this interaction component (genotype \times environment) is also studied. Accordingly, a modified concept called generalized genetical genomics was proposed by Li et al. [112], which combines both the genetic as well as carefully chosen environmental perturbations, to study the plasticity of molecular networks. This will help in understanding how a genotype responds to different environmental conditions. The utility of this approach was demonstrated in *Arabidopsis* by identifying

G × E interactions in the metabolism of germinating seeds [113]. Although these studies offer lots of information, the number of such studies in crop plants are not many and may be due to the cost associated with such experiments [113].

2.10 Identification of QTLs Using Multiparental Mapping Populations: Joint Linkage-Association Mapping

Generally, QTL mapping is carried out using a biparental mapping population for which parental genotypes exhibit contrasting phenotypes for the trait of interest. However, it is well recognized that such a mapping population will segregate for only those alleles/QTLs for which the parental genotypes differ. This leaves out many important QTLs that are controlling the trait but are not detected just because the parental genotypes do not segregate for them. Therefore, another important approach based on the principle of LD called association mapping (AM), also called genome wide association studies (GWAS), was suggested. Large numbers of studies involving AM have been published in different crop plants and are beyond the scope of this chapter. For further details, readers are referred to another chapter on this aspect in this book as well as detailed reviews [114–116]. It was also realized that linkage-based interval mapping and LD-based AM have their own advantages and limitations when used independently and therefore it was proposed to integrate these two approaches into one approach called joint linkage-association mapping (JLAM) [117]. This type of analysis has been facilitated by the availability of next-generation multiparental mapping populations like Multi-parent Advanced Generation Intercross (MAGIC) populations, Nested Association Mapping (NAM) populations, Multiline Cross Inbred Lines (MCILs), and Recombinant Inbred Advanced Intercross Lines (RIAILs) [2, 118].

These populations have been developed in many important crops including wheat, rice, maize, chickpea, pigeonpea, peanut, barley, oat, and tomato (for details, see review by [119, 120]). Although it may not be feasible to develop multiparental populations in all crops, alternatively one can perform JLAM using a number of biparental populations as well as an association-mapping population genotyped with a common set of markers. Several variants of JLAM were later also proposed including that for the analysis of multi-trait data [121–123]. The utility of JLAM was shown by Lu et al. [124] in maize. Using the NAM population, they identified 18 new QTLs and candidate genes for drought tolerance, which were earlier not identified by either of the two methods individually. Recently, in rapeseed, this method has identified two major pleiotropic QTLs for seed weight and siliques length [125]. Another advantage of using JLAM is that it can effectively address the issue of rare alleles, which is a matter of concern in any AM study [114]. Looking into its important features, this method will be used for a long time in many more crops.

2.11 Quantitative Resistance Loci (QRLs) Governing Quantitative Disease Resistance (QDR)

It is now well recognized that disease resistance in crop plants is quantitative in nature, involving major as well as minor QTLs. Accordingly, they are described either as R genes (having major effect) or quantitative resistance loci (QRL), which governs quantitative disease resistance or QDR in crop plants [126]. While dealing with QRL, the data on QDR are analyzed in the same way as that of any QTL analysis experiment for any morphologic or agronomic trait. It is therefore unnecessary to make a distinction between QRLs and QTLs. This is also evident from the fact that in several earlier studies involving QDR, the term QTL was used instead of QRL. In the last few years, large numbers of these so-called QRLs have already been identified in different crop plants including cereals and legumes, which subsequently led to map-based cloning of some of these QRLs. A partial list of such cloned QRLs in cereals is available in Gupta et al. [2].

In recent years advances in whole genome sequencing accompanied by the availability of high-throughput marker approaches like GBS has brought down the cost of genotyping drastically. These advances in genotyping technologies, if accompanied with precise and high-throughput phenotyping for QDR, will definitely facilitate the elucidation of complex forms of disease resistance and QRLs associated with them in crop plants [127–129]. It is expected that the knowledge gained from detailed understanding of QDR and that of associated QRLs will help in breeding varieties for disease resistance in crop plants in coming years. An optimal strategy is therefore needed to effectively and efficiently use the identified QRLs in breeding programs aimed at disease resistance [128, 129].

Some of the earlier successful examples of MAS for QRLs include: (1) MAS for single QRL for Fusarium head blight (FHB) in wheat [130], leaf rust in barley [131], white mold in common bean [132]; (2) multiple QRLs (pyramiding or stacking) for stripe rust in barley [133], common bacterial blight (CBB) in common bean [134], FHB in wheat ([135]; for a review, see Miedaner and Korzun [136]), root and stem rot in pepper [137]; and (3) QRLs plus qualitative resistance genes for stripe rust in barley [138], bean golden mosaic virus (BGMV) in common bean [139], potato virus Y in pepper [140], and many others.

2.12 Discovery and Introgression of Useful QTLs from Wild-Type or Unadapted Germplasm: Advanced Backcross QTL Analysis

One of the reasons often attributed to the limited use of identified QTLs in crop improvement programs is that QTL identification and varietal development are considered as separate activities. In order to deal with this issue and to harness the potential of the wild/unadapted germplasm in breeding programs, Tanksley and

Nelson [141], while working on tomato, proposed a novel method of QTL mapping called advanced backcross QTL (AB-QTL) analysis. The important feature of this method is that one can simultaneously detect and transfer useful QTLs from the wild/unadapted relatives to a popular cultivar. The backcross population (BC₂, BC₃) is developed from a cross between the superior cultivar and a wild species carrying the desirable trait, and molecular markers are used to monitor the transfer of desirable QTLs.

It is a means of reducing the number of donor parent alleles present in any given backcross inbred line. The reason for delaying QTL analysis until an advanced generation like BC₂, BC₃ is that it allows the phenotypic selection to reduce the frequency of deleterious alleles and at the same time favorable donor alleles at QTL can be more easily recognized. Since its demonstrated success in tomato, it has been used in several crops including wheat, barley, and rice for the transfer of desirable QTLs for a variety of traits from the wild/unadapted germplasm. Details of these studies are readily available in several reviews and book chapters. In recent years, its application has been seen in barley for proline accumulation and leaf wilting under drought stress conditions [142]; in rice for salinity tolerance [143], grain shape [144], and reproductive stage drought resistance [145]; and in peanut for resistance to root knot nematode [146]. Having practical significance in breeding programs, this method is going to be used for a long time.

3 Factors Affecting Results of QTL Mapping in Plants

Several factors that influence the results of any QTL-mapping experiment have been widely discussed in the literature either using computer simulations or empirical data (e.g., [8, 147–149]). Important factors amongst them are trait heritability, nature and size of mapping population, number of markers, and method of analysis (Table 2). All these factors are related to each other. For example, a mapping population of an average size of $n = 200$ will yield a low-density linkage map, which in turn will limit the precision and resolution of the QTL so identified. The end result will be that the estimates of QTL effects will be biased as QTLs with small effects will not be identified and those that are closely linked will not be separated. These factors are discussed in more detail in the following sections. There are other issues that should be considered before initiating the QTL-mapping experiment, and which have been discussed in greater detail by Wurschum et al. [150].

3.1 Heritability of the Trait

It is a well-known fact that the majority of the quantitative traits exhibit poor heritability, which makes it difficult to detect a minor effect QTL with a smaller

Table 2 Factors influencing results of QTL mapping using biparental populations

Factor	Details
Size of mapping population	More the number of individuals in the population, more accurate will be the linkage map and more accuracy in the QTL results; chances of detecting QTL with minor effect is high with larger population size
Nature of mapping population	$F2 < BC < DH \leq RILs$
Density and coverage of markers in the linkage map	More the markers on the map, less the interval distance between two markers and more accuracy in the results
Statistical method used	$SMA < SIM < CIM < ICIM \leq BIM$
Heritability of the trait	More the heritability of the trait, more the chances of QTL detection
Significance criteria used	More false positives with arbitrary significance criteria; robustness and accuracy increases with permutation test and threshold values
Effect of environment	If the effect size of the QTL is small, it may not be detected in all the environments
Experimental error	Precision in phenotyping is crucial; errors in scoring of genotypic data as well as missing marker data can affect the order of markers on the linkage map and can affect the estimated QTL location

population size and limited number of markers. Another issue with low heritability traits in QTL mapping is that QTL effects are always overestimated. This has been demonstrated empirically as well as by using simulations in several studies. Although heritability of the trait cannot be increased, scoring of the data in dynamic fashion wherever possible can help in identifying the correct stage of crop growth where heritability for the given trait is highest. This can also help in identifying novel loci that are specific to the growth stage and often escape detection. Similarly, the mapping population can also be evaluated at different locations and over the years for the trait of interest to resolve location and year effects.

3.2 *Size and Nature of Mapping Population*

Often, small mapping populations are used in linkage mapping experiments. Although one can develop a framework linkage map with smaller populations, it may not be suitable for QTL mapping. Therefore, the use of larger populations has always been appreciated for bringing precision in the QTL mapping studies. It has been shown that with a population size of >200 , methods like ICIM achieve unbiased estimations of QTL position and effect. On the contrary, when using a smaller population size, there is a tendency for the QTL to be located towards the center with overestimated QTL effects [148]. Earlier also it was shown that statistical power, QTL effect estimates, and precision of QTL localization can be improved from larger populations [147, 151, 152]. Therefore, sufficiently large

populations are needed for QTL mapping studies [29]. However, population size cannot be arbitrarily increased due to increasing costs associated with phenotyping all the lines. This issue can be overcome to some extent by using a large number of markers and high-density marker maps that can increase the precision of QTL mapping.

3.3 Number of Markers in the Linkage Map

The recent advances in cost-effective high-throughput genotyping techniques have made it possible to generate thousands of data points in several crops. These advances are also being effectively utilized in several GWAS and GS experiments. However, in the majority of the earlier studies on QTL mapping, linkage maps were developed using a rather limited number of markers. Using computer simulations, it was earlier shown that a marker density of 10–20 cm is sufficient for precise QTL detection and that there is no added advantage from higher marker densities [147, 153]. It is therefore often debated whether the biparental QTL mapping studies would benefit from high-density maps. Contrary to this, later it was shown that high-density maps could increase the probability and precision of QTL detection between two recombination breakpoints and tightly linked markers could be identified [154–156]. Moreover, two tightly linked QTLs can also be separated using high-density maps [148]. However, in a recent study based on a computer simulation as well as on experimental data of DH populations in maize, it was shown that high-density maps neither improved the QTL detection power nor the predictive power for the proportion of genotypic variance explained [157]. Furthermore, they observed that the precision of QTL localization, the precision of effect estimates for small- and medium-sized QTLs, as well as the power to resolve closely-linked QTLs profited from an increase in marker density from 5 to 1 cM. However, from an MAS point of view, precise estimates of QTL effects are more desirable and these relevant parameters may outweigh the higher costs of high-density genotyping [157].

3.4 Method of Analysis

Different methods of QTL mapping have been discussed in the earlier sections. The choice of method for QTL analysis also influences the outcome of the study. For example, ICIM has been found to be more powerful in separating tightly linked QTLs than the commonly used IM [148]. As has been discussed, the importance of interacting QTLs (QQ, QE, and QQE) cannot be underestimated. Therefore, while conducting any QTL analysis, it is necessary to choose the appropriate method that will not only identify main effect QTLs, but also different interactions with higher precision.

4 Computer Programs for QTL Analysis

Over the years, several QTL mapping approaches have been proposed, making it possible to identify thousands of marker-trait associations in crop plants. Credit for these studies also goes partly to the availability of different computer programs that facilitated these studies in a rapid manner. Since the development of the popular computer program MAPMAKER/QTL [158], large numbers of such programs are now available that can be efficiently used for the identification of QTLs using either biparental QTL mapping or association mapping. The majority of these programs are available free of cost. In recent years a shift has also been seen from the use of standalone programs to open-source environments like R. It can run on a variety of platforms and has the same ability as statistical computing and graphics (<http://www.r-project.org/>). A comprehensive, though not exhaustive, list of different types of software that can perform QTL analysis, along with their features, are given in Table 3. Similarly, a detailed list of computer programs available for AM is given in Gupta et al. [114].

5 Conclusion and Outlook

During the last two decades or more, significant progress has been witnessed in the studies involving complex quantitative traits in crop plants. This has been facilitated by the availability of the cost-effective high-throughput genotyping techniques as well as the constantly improving area of statistical genomics. Several of the identified QTLs for various traits have been, and are being, successfully used in the crop improvement programs following MAS. Starting from SMA and SIM to ICIM, and more recently BIM, QTL-mapping approaches have evolved over the years. These advances not only improved the understanding and precision of the QTL-mapping results but also the outcome of the MAS program. The increasing emphasis on the identification of interacting QTLs (QQ, QE, and QQE) has also provided a new dimension to the traditional QTL mapping studies. With growing interest in the area of genetical genomics involving eQTL, pQTL, and mQTL, coupled with generalized genetical genomics, it is expected that a better understanding about the biosynthetic pathways underlying complex traits will be gained.

In the future, the approaches of biparental QTL mapping as well as of AM/GWAS, either performed independently or in combination, will be used in many more crops using the recent advances in genomics. Methods like JLAM have the ability to harvest the benefits of both the approaches together as has been successfully demonstrated in maize [150, 181]. Similarly, the recent advances in the area of GS will address the issue of minor QTLs by way of considering the effects of all the markers simultaneously. Thus, it is evident that the progress made in the area of QTL mapping is huge and will be further benefited by recent advances

Table 3 List of computer programs available for QTL analysis

Program	Important features	Reference
SAS program	ANOVA	Knapp and Bridges [159]
MAPMAKER/ QTL	SIM, DOS based	Lincoln et al. [158]
MQTL	CIM	Tinker and Mather [160]
PLABQTL	SIM, CIM, Epistatic QTL	Utz and Melchinger [161]
QGene	SIM, CIM	Nelson [162]
SOLAR		Almasy and Blangero [163]
Multimapper	BIM	Sillanpaa and Arjas [164]
BQTL	Bayesian estimation, IM, CIM, runs on R	Berry [165]
MultiQTL	SIM, MIM	www.multiqtl.com
MapManager QTX	SIM, CIM	Manly and Olson [166]
QTL network	CIM, Epistatic QTL	Wang et al. [30]
Pseudomarker	Analysis of eQTL	Sen and Churchill [167]
QTL Express	SIM, CIM	Seaton et al. [168]
R/qtl	SIM, CIM, Epistatic QTL	Broman et al. [169]
BioMercator	Meta-analysis	Arcade et al. [96]
GridQTL	Linkage-Disequilibrium-Linkage-Analysis (LDLA) tool, epistasis	Seaton et al. [170]
Genotype matrix mapping	SIM, CIM, Epistatic QTL	Isobe et al. [171]
IciMapping	ICIM, epistasis	Li et al. [28]
MetaQTL	Meta-analysis	Veyrieras et al. [97]
QTLBIM	Map multiple interacting QTL, can handle continuous, binary and ordinal traits, R based	Yandell et al. [172]
FlexQTL	Single bi-parental mapping population up to complex multi-generations pedigrees, Bayesian analysis	Bink et al. [173]
QTLMap	Linkage analysis and linkage disequilibrium linkage analysis (LDLA); eQTL, single and multiple trait analysis	http://www.inra.fr/qtlmap
MAPQTL 6	SIM, CIM	van Ooigen [174]
QTLMiner	Discovery of candidate gene within a QTL region	Alberts and Schughart [175]
MapDisto	Linkage mapping; ANOVA	Lorieux [176]

(continued)

Table 3 (continued)

Program	Important features	Reference
Windows QTL Cartographer	SIM, CIM, MIM, multi-trait IM, BIM, Ordinal trait	Wang et al. [36]
MAPfastR	QTL mapping from inbred and outbred line-crosses; epistatic interactions, R based	Nelson et al. [177]
Dslice	Dependency detection between a categorical variable and a continuous variable, R based	Ye et al. [178]
EBEN	Multiple QTL mapping, Bayesian mapping, R based	Huang et al. [119, 120]
FastQTL	cis-QTL mapping strategy	Ongen et al. [179]
Solarius	Linkage and association mapping, R based	Ziyatdinov et al. [180]

in computational tools. The success will translate into the crop-improvement programs of the future.

References

1. Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8:552–560
2. Gupta PK, Kulwal PL, Mir RR (2013) QTL mapping: methodology and applications in cereal breeding. In: Gupta PK, Varshney RK (eds) *Cereal genomics II*. Springer, Dordrecht, pp 275–318
3. Kulwal PL, Thudi M, Varshney RK (2012) Genomics interventions in crop breeding for sustainable agriculture. In: Meyers RA (ed) *Encyclopedia of sustainability science and technology*, vol I. Springer, New York, pp 2527–2540
4. Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. *Trends Plant Sci* 10:297–304
5. Wang M, Wang S, Xia G (2015) From genome to gene: a new epoch for wheat research? *Trends Plant Sci* 20:380–387
6. Asins M (2002) Present and future of quantitative trait locus analysis in plant breeding. *Plant Breed* 121:281–291
7. Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* 142:169–196
8. Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* 3:43–52
9. Frommlet F, Bogdan M, Ramsey D (2016) *Statistical methods of QTL mapping for experimental populations. Phenotypes and genotypes*. Springer, London, pp 73–104
10. Gupta PK, Kulwal PL (2006) Methods of QTL analysis in crop plants: present status and future prospects. In: Trivedi PC (ed) *Biotechnology and biology of plants*. Avishkar Publishers, Jaipur, pp 1–23
11. Hackett CA (2002) Statistical methods of QTL mapping in cereals. *Plant Mol Biol* 48:585–599
12. Mackay TFC (2001) The genetic architecture of quantitative traits. *Annu Rev Genet* 33:303–339

13. Mauricio R (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nat Rev Genet* 2:370–381
14. Tanksley SD (1993) Mapping polygenes. *Annu Rev Genet* 27:205–233
15. Xu Y (1997) Quantitative trait loci: separating, pyramiding, and cloning. *Plant Breed Rev* 15:85–139
16. Michelmore WR, Paran I, Kesseli RV (1991) Identification of marker linked to disease resistance genes by bulked segregant analysis, a rapid method to detect the markers in specific genetic region by using the segregating population. *Proc Natl Acad Sci U S A* 88:9828–9832
17. Vikram P, Swamy BM, Dixit S, Ahmed H, Cruz MS, Singh AK, Ye G, Kumar A (2012) Bulk segregant analysis: “an effective approach for mapping consistent-effect drought grain yield QTLs in rice”. *Field Crops Res* 134:185–192
18. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, Innan H (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J* 74:174–183
19. Kujur A, Bajaj D, Saxena M, Tripathi S, Upadhyaya H et al (2014) An efficient and cost-effective approach for genic microsatellite marker-based large scale trait association mapping: identification of candidate genes for seed weight in chickpea. *Mol Breed* 34:241–265
20. Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. *Lab Anim* 30:44–52
21. Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185–199
22. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174–181
23. Jansen RC (1993) Interval mapping of multiple quantitative trait loci. *Genetics* 135:205–211
24. Zeng ZB (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci U S A* 90:10972–10976
25. Zeng ZB (1994) Precision mapping of quantitative trait loci. *Genetics* 136:1457–1468
26. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138:963–971
27. Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* 142:285–294
28. Li H, Ye G, Wang J (2007) A modified algorithm for the improvement of composite interval mapping. *Genetics* 175:361–374
29. Holland JB (2001) Epistasis and plant breeding. *Plant Breed Rev* 21:27–92
30. Wang DL, Zhu J, Li ZK, Paterson AH (1999) Mapping QTLs with epistatic effects and QTL \times environment interactions by mixed linear model approaches. *Theor Appl Genet* 99:1255–1264
31. Pang X, Wang Z, Yap JS, Wang J, Zhu J, Bo W, Lv Y, Xu F, Zhou T, Peng S, Shen D (2013) A statistical procedure to map high-order epistasis for complex traits. *Brief Bioinform* 14:302–314
32. Jannink JL, Jansen R (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* 157:445–454
33. Kulwal PL, Singh R, Balyan HS, Gupta PK (2004) Genetic basis of pre-harvest sprouting tolerance using single-locus and two-locus QTL analyses in bread wheat. *Funct Integr Genomics* 4:94–101
34. Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. *Genetics* 152:1203–1216
35. Laurie C, Wang S, Carlini-Garcia LA, Zeng Z-B (2014) Mapping epistatic quantitative trait loci. *BMC Genet* 15:112
36. Wang S, Basten CJ, Zeng Z-B (2012) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh. <http://statgen.ncsu.edu/qtlcart/WQTLCart.htm>

37. Huang A, Xu S, Cai X (2014) Whole-genome quantitative trait locus mapping reveals major role of epistasis on yield of rice. *PLoS One* 9:e87330
38. Sandhu N, Singh A, Dixit S, Cruz MT, Maturan PC, Jain RK, Kumar A (2014) Identification and mapping of stable QTL with main and epistasis effect on rice grain yield under upland drought stress. *BMC Genet* 15:63
39. Kulwal PL, Kumar N, Kumar A, Gupta RK, Balyan HS, Gupta PK (2005) Gene networks in hexaploid wheat: interacting quantitative trait loci for grain protein content. *Funct Integr Genomics* 5:254–259
40. Kumar N, Kulwal PL, Balyan HS, Gupta PK (2007) QTL analysis for yield and yield contributing traits in two mapping populations of bread wheat. *Mol Breed* 19:163–177
41. Xing W, Zhao H, Zou D (2014) Detection of main-effect and epistatic QTL for yield-related traits in rice under drought stress and normal conditions. *Can J Plant Sci* 94:633–641
42. Berger DK, Carstens M, Korsman JN, Middleton F, Kloppers FJ, Tongoona P, Myburg AA (2014) Mapping QTL conferring resistance in maize to gray leaf spot disease caused by *Cercospora zeina*. *BMC Genet* 15:60
43. Ku LX, Sun ZH, Wang CL, Zhang J, Zhao RF, Liu HY, Tai GQ, Chen YH (2012) QTL mapping and epistasis analysis of brace root traits in maize. *Mol Breed* 30:697–708
44. Bocianowski J (2013) Epistasis interaction of QTL effects as a genetic parameter influencing estimation of the genetic additive effect. *Genet Mol Biol* 36:93–100
45. Bocianowski J (2014) Estimation of epistasis in doubled haploid barley populations considering interactions between all possible marker pairs. *Euphytica* 196:105–115
46. Jiang C, Zeng Z-B (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140:1111–1127
47. Korol AB, Ronin YI, Kirzhner VM (1995) Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* 140:1137–1147
48. Korol AB, Ronin YI, Nevo E, Hays PM (1998) Multi-interval mapping of correlated trait complexes. *Heredity* 80:273–284
49. Kulwal PL, Roy JK, Balyan HS, Gupta PK (2003) QTL analysis for growth and leaf characters in bread wheat. *Plant Sci* 164:267–277
50. Apotikar DB, Venkateswarlu D, Ghorade RB, Wadaskar RM, Patil JV, Kulwal PL (2011) Mapping of shoot fly tolerance loci in sorghum using SSR markers. *J Genet* 90:59–66
51. Silva LDCE, Wang S, Zeng Z-B (2012) Multiple trait multiple interval mapping of quantitative trait loci from inbred line crosses. *BMC Genet* 13:67
52. Chen Z (2013) *Statistical methods for QTL mapping*. CRC Press, Boca Raton, pp 1–308
53. Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nat Rev Genet* 5:251–261
54. Li J, Das K, Fu G, Li R, Wu R (2011) The Bayesian lasso for genome-wide association studies. *Bioinformatics* 27:516–523
55. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 10:681–690
56. Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for Bayesian model building in genomic selection. *Genetics* 191:969–987
57. Sun X, Habier D, Fernando RL, Garrick DJ, Dekkers JC (2011) Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods. *BMC Proc* 5:1
58. Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. *Genetics* 144:805–816
59. Yi N, Xu S (2000) Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* 155:1391–1403
60. Yi N, Banerjee S, Pomp D, Yandell BS (2007) Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. *Genetics* 176:1855–1864
61. Meyer da Silva A, Leandro RA, Garcia AA, de Souza AP (2013) A Bayesian approach to map QTL and to detect epistatic effects in a maize population. *Rev Bras Biom* 31:558–581

62. Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* 63:513–521
63. Yi N, Xu S, Allison DB (2003) Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* 165:867–883
64. Kopp A, Graze RM, Xu S, Carroll SB, Nuzhdin SV (2003) Quantitative trait loci responsible for variation in sexually dimorphic traits in *Drosophila melanogaster*. *Genetics* 163:771–787
65. Bauer AM, Hoti F, Von Korff M, Pillen K, Léon J, Sillanpää MJ (2009) Advanced backcross-QTL analysis in spring barley (*H. vulgare* ssp. *spontaneum*) comparing a REML versus a Bayesian model in multi-environmental field trials. *Theor Appl Genet* 119:105–112
66. Yi N, Shriner D (2008) Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses. *Heredity* 100:240–252
67. Banerjee S, Yandell BS, Yi N (2008) Bayesian quantitative trait loci mapping for multiple traits. *Genetics* 179:2275–2289
68. Xu C, Wang X, Li Z, Xu S (2009) Mapping QTL for multiple traits using Bayesian statistics. *Genet Res* 91:23–37
69. Balestre M, Von Pinho RG, de Souza Junior CL, de Sousa Bueno Filho JS (2012) Bayesian mapping of multiple traits in maize: the importance of pleiotropic effects in studying the inheritance of quantitative traits. *Theor Appl Genet* 125:479–493
70. Shoemaker JS, Painter IS, Weir BS (1999) Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet* 15:354–358
71. Wu W-R, Li W-M, Tang D-Z, Lu H-R, Worland AJ (1999) Time-related mapping of quantitative trait loci underlying tiller number in rice. *Genetics* 151:297–303
72. Wu R, Lin M (2006) Functional mapping-how to map and study the genetic architecture of dynamic complex traits. *Nat Genet* 7:229–237
73. Kulwal PL, Ishikawa G, Benscher D, Feng Z, Yu L-X, Jadhav A, Mehetre S, Sorrells ME (2012) Association mapping for pre-harvest sprouting resistance in white winter wheat. *Theor Appl Genet* 125:793–805
74. Takai T, Yoshimichi F, Tatsuhiko S, Takeshi H (2005) Time-related mapping of quantitative trait loci controlling grain-filling in rice (*Oryza sativa* L.). *J Exp Bot* 56:2107–2118
75. Cheng X, Cheng J, Huang X, Lai Y, Wang L et al (2013) Dynamic quantitative trait loci analysis of seed reserve utilization during three germination stages in rice. *PLoS One* 8: e80002
76. Osman KA, Tang B, Wang Y, Chen J, Yu F et al (2013) Dynamic QTL analysis and candidate gene mapping for waterlogging tolerance at maize seedling stage. *PLoS One* 8:e79305. <https://doi.org/10.1371/journal.pone.0079305>
77. Würschum T, Liu W, Busemeyer L, Tucker MR, Reif JC, Weissmann EA, Hahn V, Ruckelshausen A, Maurer HP (2014) Mapping dynamic QTL for plant height in triticale. *BMC Genet* 15:59
78. Li Z, Sillanpää MJ (2013) A Bayesian nonparametric approach for mapping dynamic quantitative traits. *Genetics* 194:997–1016
79. Hackett CA, Weller JI (1995) Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* 51:1252–1263
80. Xu S, Atchley WR (1996) Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* 143:1417–1424
81. Rao S, Xu S (1998) Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity* 81:214–224
82. Xu S, Xu C (2006) A multivariate model for ordinal trait analysis. *Heredity* 97:409–417
83. Li J, Wang S, Zeng ZB (2006) Multiple-interval mapping for ordinal traits. *Genetics* 173:1649–1663
84. Feng J-Y, Zhang J, Zhang W-J, Wang S-B, Han S-F et al (2013) An efficient hierarchical generalized linear mixed model for mapping QTL of ordinal traits in crop cultivars. *PLoS One* 8(4):e59541

85. Danan S, Jean-Baptiste V, Véronique L (2011) Construction of a potato consensus map and QTL meta-analysis offer new insights into the genetic architecture of late blight resistance and plant maturity traits. *BMC Plant Biol* 11:16
86. Wu XL, Hu ZL (2012) Meta-analysis of QTL mapping experiments. In: Quantitative trait loci (QTL) methods and protocols, pp 145–171
87. Chardon F, Virilon B, Moreau L, Falque M, Joets J, Decousset L, Murigneux A, Charcosset A (2004) Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics* 168:2169–2185
88. Goffinet B, Gerber S (2000) Quantitative trait loci: a meta-analysis. *Genetics* 155:463–473
89. Lanaud C, Fouet O, Clément D, Boccara M, Risterucci AM, Surujdeo-Maharaj S, Legavre T, Argout X (2009) A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L. *Mol Breed* 24:361–374
90. Rong J, Feltus EA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL et al (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics* 176:2577–2588
91. Said JI, Lin Z, Zhang X, Song M, Zhang J (2013) A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits, drought tolerance, and disease resistance in tetraploid cotton. *BMC Genomics* 14:776
92. Swamy BM, Vikram P, Dixit S, Ahmed HU, Kumar A (2011) Meta-analysis of grain yield QTL identified during agricultural drought in grasses showed consensus. *BMC Genomics* 12:319
93. Courtois B, Ahmadi N, Khawaja F, Price AH, Rami JF, Frouin J, Hamelin C, Ruiz M (2009) Rice root genetic architecture: meta-analysis from a drought QTL database. *Rice* 2:115–128
94. Zhang H, Uddin MS, Zou C, Xie C, Xu Y, Li WX (2014) Meta-analysis and candidate gene mining of low-phosphorus tolerance in maize. *J Integr Plant Biol* 56:262–270
95. Qi Z, Sun Y, Wu Q, Liu C, Hu G, Chen Q (2011) A meta-analysis of seed protein concentration QTL in soybean. *Can J Plant Sci* 91:221–230
96. Arcade A, Labourdette A, Falque M, Mangin B, Chardon F, Charcosset A, Joets J (2004) BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* 20:2324–2326
97. Veyrieras JB, Goffinet B, Charcosset A (2007) MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinformatics* 8:49
98. Joosen RVL, Ligterink W, Hilhorst HWM, Keurentjes JJB (2009) Advances in genetical genomics of plants. *Curr Genomics* 10:540–549
99. Jansen RC, Nap J-P (2001) Genetical genomics: the added value from segregation. *Trends Genet* 17:388–391
100. Breitling R, Li Y, Tesson BM, Fu J, Wu C et al (2008) Genetical genomics: spotlight on QTL hotspots. *PLoS Genet* 4:e1000232
101. Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM (2009) Defining gene and QTL networks. *Curr Opin Plant Biol* 12:241–246
102. Keurentjes JJB, Koornneef M, Vreugdenhil D (2008) Quantitative genetics in the age of omics. *Curr Opin Plant Biol* 11:123–128
103. Kliebenstein DJ (2007) Metabolomics and plant quantitative trait locus analysis—the optimum genetical genomics platform? In: Nikolau BJ, Wurtele ES (eds) Concepts in plant metabolomics. Springer, Dordrecht, pp 29–44
104. Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? *Trends Genet* 25:39–48
105. Alseekh S, Tohge S, Wendenberg R, Scossa F, Omranian N, Li J, Kleessen S, Giavalisco P, Pleban T, Mueller-Roeber B, Zamir D (2015) Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. *Plant Cell* 27:485–512
106. Eckardt NA (2008) Epistasis and genetic regulation of variation in the Arabidopsis metabolome. *Plant Cell* 20:1185–1186

107. Lisec J, Meyer RC, Steinfath M, Redestig H, Becher M, Witucka-Wall H, Fiehn O, Törjék O, Selbig J, Altmann T et al (2008) Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations. *Plant J* 53:960–972
108. Carreno-Quintero N, Acharjee A, Maliepaard C, Bachem CW, Mumm R, Bouwmeester H, Visser RG, Keurentjes JJ (2012) Untargeted metabolic quantitative trait loci analyses reveal a relationship between primary metabolism and potato tuber quality. *Plant Physiol* 158:1306–1318
109. Feng J, Long Y, Shi L, Shi J, Barker G, Meng J (2012) Characterization of metabolite quantitative trait loci and metabolic networks that control glucosinolate concentration in the seeds and leaves of *Brassica napus*. *New Phytol* 193:96–108
110. Hill CB, Taylor JD, 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 Sci* 233:143–154
111. Luo J (2015) Metabolite-based genome-wide association studies in plants. *Curr Opin Plant Biol* 24:31–38
112. Li Y, Breitling R, Jansen RC (2008) Generalizing genetical genomics: getting added value from environmental perturbation. *Trends Genet* 24:518–524
113. Joosen RV, Arends D, Li Y, Willems LA, Keurentjes JJ, Ligterink W, Jansen RC, Hilhorst HW (2013) Identifying genotype-by-environment interactions in the metabolism of germinating *Arabidopsis* seeds using generalized genetical genomics. *Plant Physiol* 162:553–566
114. Gupta PK, Kulwal PL, Jaiswal V (2014) Association mapping in crop plants: opportunities and challenges. *Adv Genet* 85:109–147
115. Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol* 57:461–485
116. Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. *Plant Genome* 1:5–20
117. Wu R, Zeng ZB (2001) Joint linkage and linkage disequilibrium mapping in natural populations. *Genetics* 157:899–909
118. Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr Opin Plant Biol* 11:215–221
119. Huang BE, Verbyla KL, Verbyla AP, Raghavan C, Singh VK, Gaur P, Leung H, Varshney RK, Cavanagh CR (2015) MAGIC populations in crops: current status and future prospects. *Theor Appl Genet* 128:999–1017
120. Huang A, Xu S, Cai X (2015) Empirical Bayesian elastic net for multiple quantitative trait locus mapping. *Heredity* 114:107–115
121. Meuwissen TH, Goddard ME (2004) Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet Sel Evol* 36:261–279
122. Stich B, Piepho H-P, Schulz B, Melchinger AE (2008) Multi-trait association mapping in sugar beet (*Beta vulgaris* L.) *Theor Appl Genet* 117:947–954
123. Wu R, Chang-Xing M, George C (2002) Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. *Genetics* 160:779–792
124. Lu Y, Zhang S, Shah T, Xie C, Hao Z, Li X, Farkhari M, Ribaut J-M, Cao M, Rong T, Xu Y (2010) Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proc Natl Acad Sci U S A* 107:19585–19590
125. Li N, Shi J, Wang X, Liu G, Wang H (2014) A combined linkage and regional association mapping validation and fine mapping of two major pleiotropic QTLs for seed weight and silique length in rapeseed (*Brassica napus* L.) *BMC Plant Biol* 14:114
126. Young ND (1996) QTL mapping and quantitative disease resistance in plants. *Annu Rev Phytopathol* 34:479–501
127. Kou Y, Wang S (2010) Broad-spectrum and durability: understanding of quantitative disease resistance. *Curr Opin Plant Biol* 13:181–185

128. Poland JA, Balint-Kurti PJ, Wisser RJ, Pratt RC, Nelson RJ (2009) Shades of gray: the world of quantitative disease resistance. *Trends Plant Sci* 11:21–29
129. St. Clair DA (2010) Quantitative disease resistance and quantitative resistance loci in breeding. *Annu Rev Phytopathol* 48:247–268
130. Pumphrey MO, Bernardo R, Anderson JA (2007) Validating the *Fhbl* QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. *Crop Sci* 47:200–206
131. Marcel TC, Aghnoum R, Durand J, Varshney RK, Niks RE (2007) Dissection of the barley 2L1.0 region carrying the ‘Laevigatum’ quantitative resistance gene to leaf rust using near-isogenic lines (NIL) and subNIL. *Mol Plant Microbe Interact* 20:1604–1615
132. Miklas PN (2007) Marker-assisted backcrossing QTL for partial resistance to Sclerotinia white mold in dry bean. *Crop Sci* 47:935–942
133. Toojinda T, Baird E, Booth A, Broers L, Hayes P et al (1998) Introgression of quantitative trait loci (QTLs) determining stripe rust resistance in barley: an example of marker-assisted line development. *Theor Appl Genet* 96:123–131
134. Mutlu N, Miklas P, Reiser J, Coyne D (2005) Backcross breeding for improved resistance to common bacterial blight in pinto bean (*Phaseolus vulgaris* L.). *Plant Breed* 124:282–288
135. Wilde F, Schon CC, Korzun V, Ebmeyer E, Schmolke M et al (2008) Marker-based introduction of three quantitative-trait loci conferring resistance to Fusarium head blight into an independent elite winter wheat breeding population. *Theor Appl Genet* 117:29–35
136. Miedaner T, Korzun V (2012) Marker-assisted selection for disease resistance in wheat and barley breeding. *Phytopathology* 102:560–566
137. Thabuis A, Palloix A, Servin B, Daubeze A-M, Signoret P et al (2004) Marker-assisted introgression of 4 *Phytophthora capsici* resistance QTL alleles into a bell pepper line: validation of additive and epistatic effects. *Mol Breed* 14:9–20
138. Castro AJ, Capettini F, Corey AE, Filichkina T, Hayes PM et al (2003) Mapping and pyramiding of qualitative and quantitative resistance to stripe rust in barley. *Theor Appl Genet* 107:922–930
139. Miklas PN, Kelly JD, Beebe SE, Blair MW (2006) Common bean breeding for resistance against biotic and abiotic stresses: from classical to MAS breeding. *Euphytica* 147:105–131
140. Palloix A, Ayme V, Moury B (2009) Durability of plant major resistance genes to pathogens depends on the genetic background, experimental evidence and consequences for breeding strategies. *New Phytol* 183:90–99
141. Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor Appl Genet* 92:191–203
142. Sayed MA, Schumann H, Pillen K, Naz AA, Léon J (2012) AB-QTL analysis reveals new alleles associated to proline accumulation and leaf wilting under drought stress conditions in barley (*Hordeum vulgare* L.) *BMC Genet* 13:61
143. Chai L, Zhang J, Pan XB, Zhang F, Zheng TQ, Zhao XQ, Wang WS, Jauhar A, Xu JL, Li ZK (2014) Advanced backcross QTL analysis for the whole plant growth duration salt tolerance in rice (*Oryza sativa* L.). *J Integ Agric* 13:1609–1620
144. Nagata K, Ando T, Nonoue Y, Mizubayashi T, Kitazawa N, Shomura A, Matsubara K, Ono N, Mizobuchi R, Shibaya T, Ogiso-Tanaka E (2015) Advanced backcross QTL analysis reveals complicated genetic control of rice grain shape in a *japonica* × *indica* cross. *Breed Sci* 65:308–318
145. Sellamuthu R, Ranganathan C, Serraj R (2015) Mapping QTLs for reproductive-stage drought resistance traits using an advanced backcross population in upland rice. *Crop Sci* 55:1524–1536
146. Burow MD, Starr JL, Park C-H, Simpson CE, Paterson AH (2014) Introgression of homeologous quantitative trait loci (QTLs) for resistance to the root-knot nematode [*Meloidogyne arenaria* (Neal) Chitwood] in an advanced backcross-QTL population of peanut (*Arachis hypogaea* L.) *Mol Breed* 34:393–406

147. Darvasi A, Weinreb A, Minke V, Weller JI, Soller M (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* 134:943–951
148. Li H, Hearne S, Banziger M, Li Z, Wang J (2010) Statistical properties of QTL linkage mapping in biparental genetic populations. *Heredity* 105:257–267
149. Utz HF, Melchinger AE, Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* 154:1839–1849
150. Wurschum T (2012) Mapping QTL for agronomic traits in breeding populations. *Theor Appl Genet* 125:201–210
151. Beavis WD (1998) QTL analyses: power, precision, and accuracy. In: Paterson AH (ed) *Molecular dissection of complex traits*. CRC Press, New York, pp 145–162
152. Vales MI, Schön CC, Capettini F, Chen XM, Corey AE, Mather DE, Mundt CC, Richardson KL, Sandoval-Islas JS, Utz HF, Hayes PM (2005) Effect of population size on the estimation of QTL: a test using resistance to barley stripe rust. *Theor Appl Genet* 111:1260–1270
153. Piepho HP (2000) A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics* 156:2043–2050
154. Almeida GD, Makumbi D, Magorokosho C, Nair S, Borém A, Ribaut JM, Bänziger M, Prasanna BM, Crossa J, Babu R (2012) QTL mapping in three tropical maize populations reveals a set of constitutive and adaptive genomic regions for drought tolerance. *Theor Appl Genet* 126:583–600
155. Shi LY, Hao ZF, Weng JF, Xie CX, Liu CL, Zhang DG, Li MS, Bai L, Li XH, Zhang SH (2011) Identification of a major quantitative trait locus for resistance to maize rough dwarf virus in a Chinese maize inbred line X178 using a linkage map based on 514 gene-derived single nucleotide polymorphisms. *Mol Breed* 30:615–625
156. Yu H, Xie W, Wang J, Xing Y, Xu C, Li X, Xiao J, Zhang Q (2011) Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One* 6:e17595
157. Stange M, Utz HF, Schrag TA, Melchinger AE, Würschum T (2013) High-density genotyping: an overkill for QTL mapping? Lessons learned from a Case study in maize and simulations. *Theor Appl Genet* 126:2563–2574
158. Lincoln S, Daly M, Lander E (1993) *Mapping genes controlling quantitative traits using MAPMAKER/QTL*. Version 1.1, 2nd edn. Whitehead Institute for Biomedical Research, Technical report
159. Knapp SJ, Bridges WC (1990) Using molecular markers to estimate quantitative trait locus parameters; power and genetic variances for unreplicated and replicated progeny. *Genetics* 126:769–777
160. Tinker NA, Mather DE (1995) MQTL: software for simplified composite interval mapping of QTL in multiple environments. *J Quant Trait Loci* 1:2
161. Utz H, Melchinger A (1996) PLABQTL: a program for composite interval mapping of QTL. *J Quant Trait Loci* 2:1
162. Nelson JC (1997) QGENE: software for marker-based genomic analysis and breeding. *Mol Breed* 3:239–245
163. Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198–1211
164. Sillanpaa MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* 148:1373–1388
165. Berry CC (1998) Computationally efficient Bayesian QTL mapping in experimental crosses. In: *ASA proceedings of the biometrics section*, pp 164–169
166. Manly KF, Olson JM (1999) Overview of QTL mapping software and introduction to map manager QTL. *Mamm Genome* 10:327–334

167. Sen Ś, Churchill GA (2001) A statistical framework for quantitative trait mapping. *Genetics* 159:371–387
168. Seaton G, Haley CS, Knott SA, Kearsley M, Visscher PM (2002) QTL express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics* 18:339–340
169. Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19:889–890
170. Seaton G, Hernandez J, Grunchev JA, White I, Allen J, De Koning DJ, Wei W, Berry D, Haley C, Knott S (2006) GridQTL: a grid portal for QTL mapping of compute intensive datasets. In: *Proceedings of the 8th world congress on genetics applied to livestock production*. Belo Horizonte
171. Isobe S, Nakaya A, Tabata S (2007) Genotype matrix mapping: searching for quantitative trait loci interactions in genetic variation in complex traits. *DNA Res* 14:217–225
172. Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, Von Smith R, Yi N (2007) R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics* 23:641–643
173. Bink MCAM, Boer MP, ter Braak CJF, Jansen J, Voorrips RE, van de Weg WE (2008) Bayesian analysis of complex traits in pedigreed plant populations. *Euphytica* 161:85–96
174. van Ooijen JW (2009) MapQTL R 6, software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma BV, Wageningen
175. Alberts R, Schughart K (2010) QTLminer: identifying genes regulating quantitative traits. *BMC Bioinformatics* 11:516
176. Lorieux M (2012) MapDisto: fast and efficient computation of genetic linkage maps. *Mol Breed* 30:1231–1235
177. Nelson RM, Nettelblad C, Pettersson ME, Shen X, Crooks L, Besnier F, Álvarez-Castro JM, Rönnegård L, Ek W, Sheng Z, Kierczak M (2013) MAPfastR: quantitative trait loci mapping in outbred line crosses. *G3: Genes Genom Genet* 3:2147–2149
178. Ye C, Jiang B, Zhang X, Liu JS (2015) dslice: an R package for nonparametric testing of associations with application in QTL and gene set analysis. *Bioinformatics* 31:1842–1844
179. Ongen H, Buil A, Brown AA, Dermizakis ET, Delaneau O (2015) Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32(10):1479–1485
180. Ziyatdinov A, Brunel H, Martinez-Perez A, Buil A, Perera A, Soria JM (2016) solarius: an R interface to SOLAR for variance component analysis in pedigrees. *Bioinformatics* 32(12):1901–1902
181. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2202

Trait Mapping Approaches Through Association Analysis in Plants



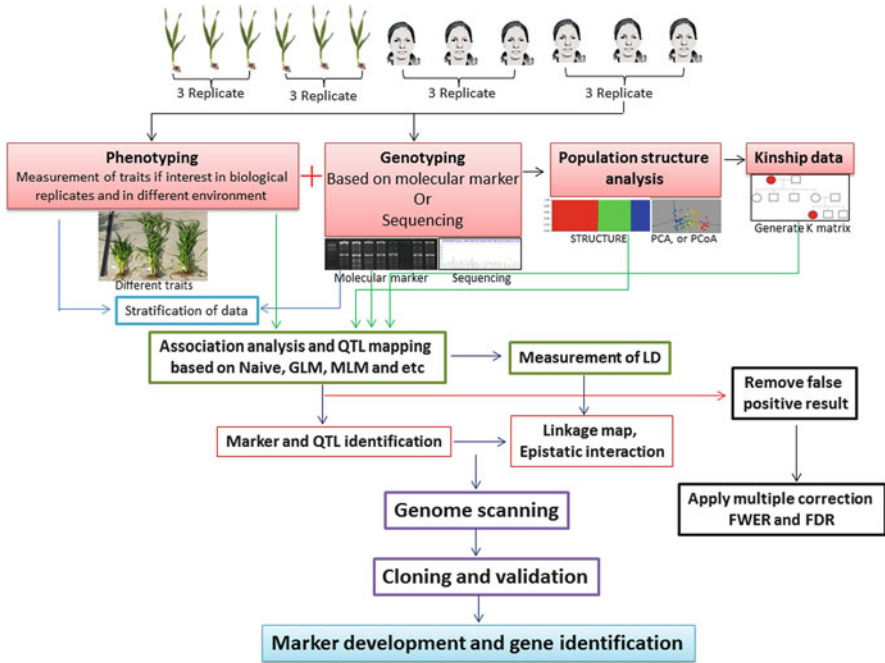
M. Saba Rahim, Himanshu Sharma, Afsana Parveen, and Joy K. Roy

Abstract Previously, association mapping (AM) methodology was used to unravel genetic complications in animal science by measuring the complex traits for candidate and non-candidate genes. Nowadays, this statistical approach is widely used to clarify the complexity in plant breeding program-based genome-wide breeding strategies, marker development, and diversity analysis. This chapter is particularly focused on methodologies with limitations and provides an overview of AM models and software used up to now. Association or linkage disequilibrium mapping has become a very popular method for discovering candidate and non-candidate genes and confirmation of quantitative trait loci (QTL) on various parts of the genome and in marker-assisted selection for breeding. Previously, various QTL investigations were carried out for different plants exclusively by linkage mapping. To help to understand the basics of modern molecular genetic techniques, in this chapter we summarize previous studies done on different crops. AM offers high-resolution power when there is large genotypic diversity and low linkage disequilibrium (LD) for the germplasm being investigated. The benefits of AM, compared with traditional QTL mapping, include a relatively detailed mapping resolution and a far less time-consuming approach since no mapping populations need to be generated. The advancements in genotyping and computational techniques have encouraged the use of AM. AM provides a fascinating approach for genetic investigation of QTLs, due to its resolution and the possibility to study the various genomic areas at the same time without construction of mapping populations. In this chapter we also discuss the advantages and disadvantages of AM, especially in the dicotyledonous crops Fabaceae and Solanaceae, with various genome-size reproductive strategies (clonal vs. sexual), and statistical models. The main objective of this chapter is to highlight the uses of association genetics in major and minor crop species that have

M. Saba Rahim, H. Sharma, A. Parveen, and J. K. Roy (✉)
National Agri-Food Biotechnology Institute (NABI), Mohali, India
e-mail: joykroy@nabi.res.in

trouble being analyzed for dissection of complex traits by identification of the factor responsible for controlling the effect of trait.

Graphical Abstract



Keywords Association mapping (AM), Linkage disequilibrium (LD), Marker-assisted selection (MAS), Quantitative trait loci (QTLs)

Contents

1	Introduction	85
2	Trait Mapping Approaches	86
3	Objectives of Trait Mapping	87
4	Steps for Association Mapping	88
5	Advances and Scope (Methodology)	88
6	“STRUCTURE” Run Parameters (Ancestry Model)	89
6.1	Admixture Model	90
6.2	No Admixture Model	90
6.3	Linkage Model	90
7	Estimation of Sub-populations (K)	90
8	Analyzing the Results	91
8.1	Summary of “STRUCTURE” Output	91
8.2	Ancestry Estimates	92
8.3	Plots of Summary Statistics	92
8.4	Histogram Plots of <i>Fst</i> and <i>alpha</i>	93

9 Why Do Association Mapping (AM)? 93

10 Stratification of Data 94

11 Input File Required for AM Using a General Linear Model (GLM) 94

12 Input File Required for AM Using a Mixed Linear Model (MLM) 94

13 Coefficient of Kinship Data 95

14 Models Used in AM 96

15 Presentation of the Statistical Model in AM 96

16 Statistics for Phenotypic Trait and Association Analysis 96

17 Correction of “Type I” and “Type II” Errors 96

18 Model Selection for Marker-associated Trait 96

19 Application 96

20 Limitations 97

21 Conclusion 105

References 105

Abbreviations

AM	Association mapping
CV	Coefficient variance
EST	Expressed sequence tags
FDR	False discovery rate
FWER	Family-wise error rate
GLM	General linear model
GS	Genomic selection
GWAS	Genome-wide association study
LD	Linkage disequilibrium
MAS	Marker-assisted selection
MCA	Multiple correspondence analysis
MCMC	Markov chain Monte Carlo
MLM	Mixed linear model
MLMM	Multiple locus multiple marker
MTMM	Multiple trait multiple marker
PCA	Principal component analysis
QTL	Quantitative trait locus
SA	Structure analysis
SLST	Single locus single trait
SNP	Single nucleotide polymorphism

1 Introduction

Population genetics was derived from Mendel’s theory in 1900 and explains the concept of heredity in science. Further, it explains that phenotypic variation can be affected by environmental conditions [1]. Nowadays it has a great impact on agriculture in the study of evolutionary and molecular biology. The complexity of

phenotypic traits is related to segregation of alleles and the interactions between loci controlling the effects of individual traits. In modern genetics, basic statistics makes it possible to understand genetic changes and to identify the chromosome region involved. In this chapter we describe the advancements in association mapping (AM), their methodology, different statistics models, population types, traits used in plants, and limitations with a special focus on developing the understanding of marker-trait associations for the breeding community.

AM was widely used as a statistical method in animal science for high-resolution, genome-wide association analysis for several diseases such as diabetes and cancer [2], to translate the susceptibility of traits with a complete description of associated diseases [3]. In plant science, AM studies are used to identify the marker trait associations. In addition, the associated marker is used in marker-assisted breeding for phenotype selection, and in this way it is more efficient, reliable, and cost effective as compared to traditional breeding methodology [4]. Thus, AM is a strategy that applies from phenotype to genotype, localizing the chromosomal region that might contain a gene or a cluster of genes that contribute phenotypic variation. The removal of obstructions in breeding programs is required for the improvement of crops by facilitating high-resolution mapping of adapted diversification, but it is challenging to identify a locus that controls the trait of variation. AM and linkage mapping are two widely used methods to identify quantitative trait loci (QTLs) with genetically linked molecular markers, which are used for incorporating genes into cultivars via map-based cloning of the tagged gene.

AM has opened the path in agriculture for QTL analysis and marker-assisted selection (MAS). Many important traits such as crop yield, quality, abiotic resistance, disease resistance, and adaptation are due to polygenic effects measured among individuals through the action of genes and their interaction in different environmental conditions. The selection of a population is an important factor in conducting a preliminary genetic map based on association analysis. In this chapter we address the limitations and application of AM in plant science. We also detail the methods and statistics used in AM, and list complete information such as marker number and type, germplasm number and type, statistics, and software used in association and QTL mapping.

2 Trait Mapping Approaches

The basic objective of AM studies is to detect correlations between genotypes and phenotypes in a sample of individuals on the basis of linkage disequilibrium (LD) [5]. AM is an alternative of QTL mapping that does not require development of bi-parental crosses or screening generation of progeny. Thus, AM is a statistical assessment of the association between genotypes and phenotypes, and we can apply this approach to detecting QTL for traits that show variation [6]. We applied AM in crops for the identification of genetic markers sharing an association with traits. In this approach, the pre-selection of genotypes is necessary, such as linked or unlinked markers, for better elucidation of genetic linkage [7]. Several authors claim that two to four markers per chromosome are needed for candidate gene association. However, the number of chromosomes and diversity among the sample affect genotype study.

Several molecular markers such as RFLP, RAPD, AFLP, SSR and DArT, SNP, and EST have been used for AM. In the past, protein-based markers and isoenzymes were used to detect sequence differences between two individuals. Important advantages of the AM include sampling of complex or unrelated individuals in the plant population as well as human disease, marker-assisted selection in plant breeding [8], and studies of several phenotypic traits in the same population by using the same genotypic data.

An ideal sample with subtle population structure and familial relatedness, a multi-family sample, a sample with population structure, a sample with both population structure and familial relationships, and a sample with severe population structure and familial relationships determined the amenable association studies [9, 10]. The phenotypic data are dependent on traits being analyzed. The screening of more complex traits is more valuable for trait mapping. AM studies in many major crops such as rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), vegetables such as tomato (*Lycopersicon esculentum* L.), eggplant (*Solanum melongena* L.), potato (*Solanum tuberosum* L.), grasses such as sugarcane (*Saccharum officinarum* L.), Arabidopsis plant, as well as trees such as aspen (*Populus tremula* L.) and loblolly pine (*Pinus taeda* L.) have already been conducted for several traits including plant height, heading date, heading time [11], tiller number, tiller angle, flag leaf length, flag leaf width, pericarp color [12], kernel weight, kernel width, kernel area, kernel length, higher flour yield [13], grain yield, bio-ethanol production [14], tolerance to pre-harvest sprouting [15], number of spikelets/spikes, spike length, grain protein content, hardness index [16], starch, oil, moisture [17], spot blotch resistance [6], fruit weight, fruit length, fruit curvature, flesh color, plant growth habit, leaf width, leaf length [18], amino acid, organic acid, seven phenylpropanoids, and other metabolites [19] (Fig. 1 and Table 1).

3 Objectives of Trait Mapping

- AM of appropriate traits
- Evaluate the factors controlling a phenotype throughout the population
- Develop marker/s

Table 1 Molecular markers used in trait mapping

Molecular markers	Acronym
Restriction fragment length polymorphism	RFLP
Random amplified polymorphic DNA	RAPD
Short sequence repeats	SSR
Amplified fragment length polymorphism	AFLP
Single nucleotide polymorphism	SNP
Variable number tandem repeats	VNTR
Presence absence variance	PAV
Diversity arrays technology	DArT
Sequence characterized amplified region	SCAR
Allele specific associated primer	ASAP

- Design a genetic construct that shows the major difference between two varieties of a particular trait
- Identify disease carrier or resistance
- Estimation of genetic distance
- Discover and analyze genes associated with traits.

4 Steps for Association Mapping

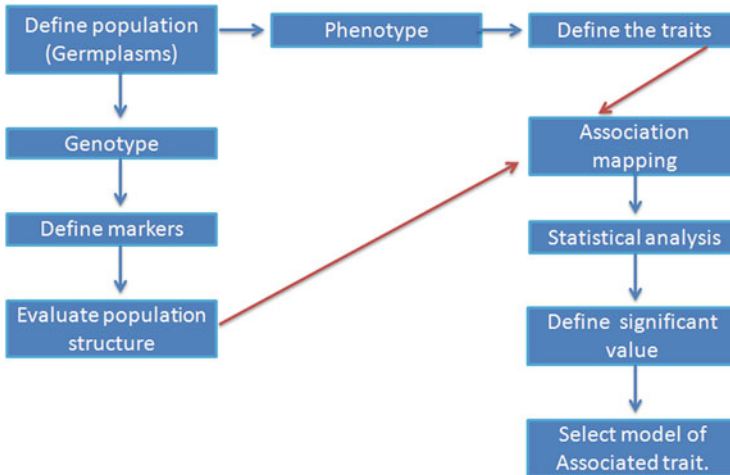


Fig. 1 Flow chart showing the steps involved in association mapping (AM)

5 Advances and Scope (Methodology)

A Bayesian approach for the inference of population structure based on markers is implemented in the computer program “STRUCTURE [22].” Several other types of software are enabled for population analysis such as FRAPP, EIGENSOFT, PLINK, and HAPMIX. The recently released StrAuto v0.3.1 is a Python-based structure software with an automated approach for linux-based computers [25]. The program has been widely used for the detection of genetic structure in sample populations for medical purposes [26, 27], assignment studies [28], population structure and hybridization analysis [29–31], migration and dispersal analysis [32–34], and also for detecting the cryptic genetic structure of natural populations [35, 36] (Fig. 2).

For 2D or 3D space, multiple correspondence analysis (MCA) and principle component analysis (PCA) is performed to observe the relative dispersion of the subpopulation. It takes less computing time than maximum likelihood estimation. PCA produces a two- or three-dimensional scatter plot of the samples in which geometric distances among samples in the plot reflect the genetic distances among

AM method: Population based marker development

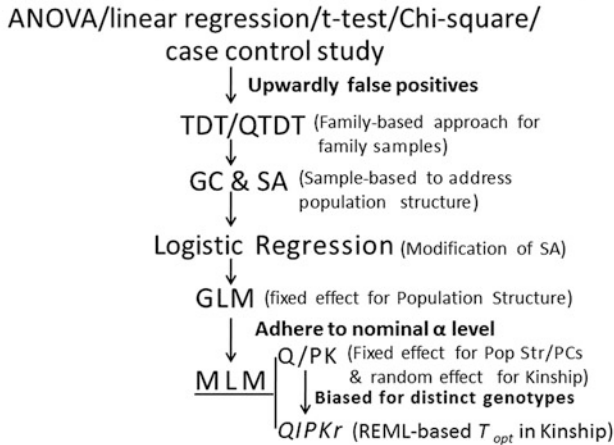


Fig. 2 Work flow to develop a population-based marker in an association-mapping (AM) panel

them with a minimum distortion and ambiguity compared to cluster analysis [37]. It can be performed only on numerical data sets that do not have missing values. Therefore, PCA is currently used more for population structure analysis and discriminate analysis, while “STRUCTURE” is widely used for the “Bayesian clustering method.” To detect the true number of clusters, we use ad hoc statistics to find ΔK based on the posterior probability in the second-order rate of change from the individual ancestry coefficient [LnP(d)] value provided by the software “STRUCTURE.” The results are sensitive to genetic markers such as AFLP and microsatellite. These microsatellite DNA markers are widely used because they are both co-dominant and highly polymorphic [38].

6 “STRUCTURE” Run Parameters (Ancestry Model)

There are lots of parameters in the default settings of *extraparam* that are mentioned in the user’s manual of “STRUCTURE” software (Pritchard et al. 2003). Among these we can choose the level of ancestry model as admixture, without admixture and linkage model, degree of admixture between population “*alpha*” to be inferred from the data, the parameter of the distribution of allelic frequencies “*lambda*,” and informativeness of the sampling location data “*r*” in *mainparam*. We set the length value of burn-in and Markov Chain Monte Carlo (MCMC); typically a burn-in of 10–100 K is more than adequate. You can choose the possible length of burn-in and MCMC, and will need to do several runs at each K.

6.1 *Admixture Model*

This is a flexible model that deals with many complexities in a population because the individuals have mixed ancestry, i.e., some fraction of the individual genome is inherited from an ancestor in the population.

6.2 *No Admixture Model*

This type of model is used when the individual originated purely from one population. The feature of this model is to analyze fully discrete populations to detect clustering.

6.3 *Linkage Model*

This is the generalized admixture model for dealing with admixture linkage disequilibrium. The detailed computations of the model are described in [39]. Briefly, we can use this model to better perform and simplify the complex of admixed populations [40].

7 **Estimation of Sub-populations (K)**

To detect the true K is an estimate of the posterior probability of the data of the given K , $\Pr(X | K)$ [22], which is called “LnP (D)” in STRUCTURE output. First, we plot the mean likelihood $L(K)$ over possible runs for each K . Second, we plot the mean difference between the successive likelihood values of K , $L'(K) = L(K) - L(K-1)$, this is the first-order rate of change. In the third step we plot the difference between the successive likelihood values of $L'(K)$, $|L''(K)| = |L'(K+1) - L'(K)|$. This corresponds to the second-order rate of change of $L(K)$ with respect to K . Finally, we estimate ΔK as the mean of the absolute values of $L''(K)$, averaged over possible runs, divided by the standard deviation of $L(K)$, $\Delta K = m(|L''(K)|) / s[L(K)]$. We find the modal value of the distribution of ΔK to be located at the real K . The graph indicates the strength of the clear peak at the true value of K [41].

Several studies carried out genomic control (GC) and structured association (SA) to overcome the effect of ambiguous structure [26]. Principle component analysis (PCA) is the best way to analyze genetic diversity and at the level of admixture population structure analysis, it is an effective way to diagnose the population structure [21, 42]. This analysis is based on correlation as well as covariance between the variables, on the basis of principle components. In PCA, Q (Membership coefficient) is replaced by a loading factor of each individual that describes the population membership of the individual.

Alternatively, we can classify the population according to the germplasm collection based on sources; they are derived from wild populations or breeding germplasm, synthetic populations, and elite germplasm [13].

8 Analyzing the Results

8.1 Summary of “STRUCTURE” Output

STRUCTURE by Pritchard, Stephens and Donnelly (2000)
And Falush, Stephens and Pritchard (2003)
Code by Pritchard, Falush and Hubisz
Version 2.3.4

Run parameters:
10 individuals
67 loci
3 populations assumed
10000 Burn-in period
100000 Reps

Estimated **Ln Prob of Data** = -9535.7
Mean value of ln likelihood = -9362.8
Variance of ln likelihood = 345.9
Mean value of **alpha** = 0.1509
Mean value of Fst_1 = 0.2685
Mean value of Fst_2 = 0.2193
Mean value of Fst_3 = 0.2080

Inferred ancestry of individuals (Q)

Label	(%Miss)	:	Inferred clusters		
1 A	(12)	:	0.239	0.449	0.311
2 B	(8)	:	0.246	0.740	0.014
3 C	(11)	:	0.347	0.640	0.013
4 D	(14)	:	0.004	0.007	0.989
5 E	(22)	:	0.291	0.029	0.681
6 F	(11)	:	0.234	0.427	0.338
7 G	(16)	:	0.989	0.007	0.004
8 H	(13)	:	0.986	0.010	0.004
9 I	(23)	:	0.980	0.007	0.013
10 J	(13)	:	0.060	0.759	0.181

There are several types of plots of ancestry estimates and plots of summary statistics. Histogram plots of *Fst* and *alpha* are shown in the text result.

8.2 Ancestry Estimates

There are two types of plots provided for the *Q* (estimated membership coefficient of individual). In these types of bar blot, each individual in the data set is represented by a single vertical line, partitioned into *K* color segments that represent the inferred cluster. Another type of plot is visualized for the *Q* into a triangle that explores the data for *K* = 3 [43] (Figs. 3 and 4).

8.3 Plots of Summary Statistics

During the course of running the software program plot, the time-series plots for each *K* that summarizes the brief period at the start of the run where the value increases up to stationary distribution at the end of burn-in (Fig. 5).

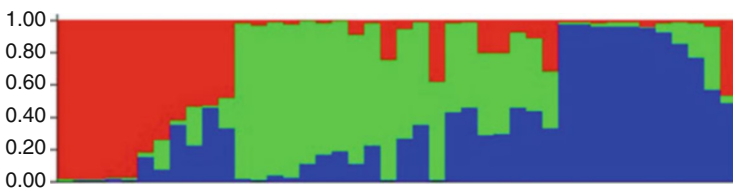


Fig. 3 The bar plot represents sub-populations arranged according to their most likely ancestry

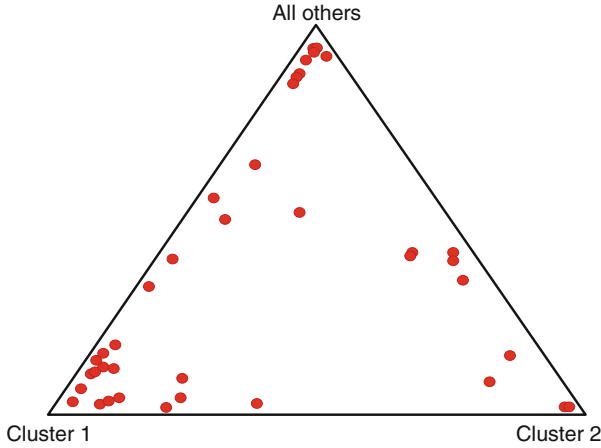


Fig. 4 Triangular plot developed by “STRUCTURE” that represents sub-populations

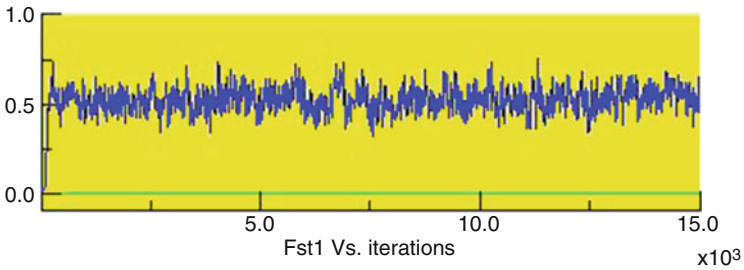


Fig. 5 Time series plot of F_{ST}

8.4 Histogram Plots of Fst and alpha

In a population structure, F_{st} is useful to examine the overall genetic divergence relative to the subpopulation within the total population.

9 Why Do Association Mapping (AM)?

- To discover the linked marker/s associated with a gene that controls the trait.
- To ascertain if the effect of a gene is either additive or dominant.
- To exploit the natural variation found in a species
- Landraces
- Cultivars from multiple programs
- Variation from regional breeding programs.

In plants and animals, AM study is the implementation of trait mapping by using genetic marker information. In this approach, the estimated membership coefficient value (Q) from the structure output is further used for structure association. The use of genetic markers to assist trait mapping is successful in marker-assisted selection (MAS), and genomic selection (GS) for breeding strategy. These population genetics studies not only allow researchers to integrate studies for need interests but also allow a deep understanding of candidate genes and dissection of related complex traits. The hypothesis of the association of genetic markers with traits is tested by different algorithms such as the mixed linear model (MLM) based on Kinship matrix (K – model), both the $K + Q$ model, and the general linear model (GLM). Based on the Q matrix, single-locus single traits (SLST), multi-locus mixed model (MLMM), and multi-trait mixed model (MTMM) have been proposed. Genome-wide association analysis (GWAS) is involved for the dissection of a large complex trait analysis. The GWAS presents the best understanding of the genetic architecture of the traits of a crop [15].

10 Stratification of Data

For the accuracy and validity of associations, several studies have applied STRAT-based stratification to improve the sample size, number of loci, and degree of divergence between populations [22]. STRAT-based stratification can also be used when two or more populations are admixed [44, 45]. Campbell et al. [46] studied and analyzed the efficacy of stratification by constructing a case-control group with the presence or absence of stratification.

11 Input File Required for AM Using a General Linear Model (GLM)

- Genotypic data (Molecular markers)
- Phenotypic data (Traits)
- Covariates (Q matrices)

12 Input File Required for AM Using a Mixed Linear Model (MLM)

This is similar to running GLM but the difference is that it requires Kinship data (K).

13 Coefficient of Kinship Data

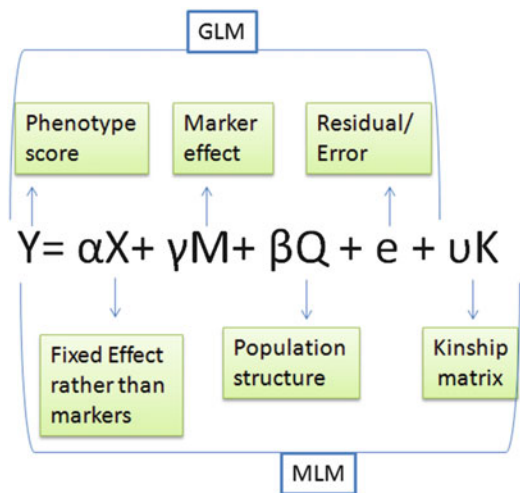
The K matrix is developed by marker data that provide more information about relatedness among individuals.

In AM analysis, an individual statistical model contains dependent variables such as trait/s data and independent variables such as marker data. In $Q + K$ models of AM, Q matrices show variables as fixed effects and K matrices show variables as random effects (Table 2 and Fig. 6).

Table 2 Summary of models used in association mapping

S. N.	Model	Description	References
1	NAIVE	Simple test of association (Kruskal-Wallis) with no correction for population structure	Thornsberry et al. [20] Yu et al. [10]
2	Q	Inferred population structure as cofactor, i.e., structured association	Price et al. [21] Pritchard et al. [22]
3	K	Mixed model without inferred population structure as cofactor	Zhao et al. [23]
4	$Q + K$	Mixed model with inferred population structure as fixed effect	
5	K^*	Same as K , but using an alternative kinship matrix based on haplotype sharing	
6	$Q + K^*$	Same as $Q + K$, but using an alternative kinship matrix based on haplotype sharing	
7	P	PCA	
8	$P + K$	Same as $Q + K$, but using P instead of Q	
9	$P + K^*$	Same as $Q + K^*$, but using P instead of Q	

Fig. 6 Statistical model defining the function used in a general linear model and mixed linear model



14 Models Used in AM

15 Presentation of the Statistical Model in AM

16 Statistics for Phenotypic Trait and Association Analysis

A model-based clustered analysis of AM was performed earlier [47]. Through descriptive statistical analysis including frequency distribution, mean value, coefficient of variability (CV), and Pearson's correlation coefficient, we can find an association between genetic information and phenotypic variation at a molecular level. Correlations based on LD are the primordial statistics of AM [48]. Gupta et al. [49] have already discussed the different factors affecting the LD, their current issues, and uses in plant sciences.

17 Correction of "Type I" and "Type II" Errors

Due to the presence of another variable or type I and II errors, AM shows confounding results or gives spurious associations. There are two multiple significance tests that are required to reduce the chance of false association, (I) Family-Wise Error Rate (FWER), and (II) False Discovery Rate (FDR). FDR is based on statistical models to remove "Type I" error [50] and "Type II" error [51], and gives the most conservative Bonferroni-corrected significance level. New approaches of FDR have also been developed to control the FWER.

18 Model Selection for Marker-associated Trait

The following two criteria were used for model selection, lowest mean of squared difference (MSD) between the observed and expected p value of all marker loci, and percentage of observation that is below the nominal level ($\alpha = 0.05$) in a p (expected) – p (observed) plot quantile–quantile plot (Q–Q plot).

19 Application

- AM is usually performed and genome type based selection of individual in plant species is applied.
- Genome-wide association analysis in different plant species.

- Comprehensive genome scans can be built through intensive sequencing and high-density genotyping.
- In breeding, several national laboratories have been able to advance the research work in marker development and marker-assisted selection through trait mapping.
- Linkage analysis and map construction.
- Dissection of gene-associated complex traits to find genes or a genomic region can move toward economically and evolutionary valuable traits for superior research.
- For parental selection, a mixed model is used to calculate the breeding values in the aid of selecting parents for crossing.
- Through this approach we can define bi-parental populations of rare alleles and emphasize the study of epistatic interactions.

20 Limitations

- AM has higher probabilities of type I and type II errors than QTL analysis. Type I error or false positives arise from unaccounted subdivisions in the sample, referred to as population structure [22].
- QTL analysis is attributed at least three factors: (1) lower correlation between markers and genes due to the decay of LD, (2) the presence/absence of alleles at different frequencies, (3) a serious multiple testing problem, which results in an extremely strict genome-wide significance threshold [52].
- The hexaploid nature of the wheat genome has introduced more difficulty for AM compare to other crops having less complex genomes.
- Due to random mating in the sampling population and some individuals being more closely related than others, some authors conduct the analysis within sub-populations [53, 54] to avoid this problem.
- When the mode of ΔK at the true K was absent, it was either because sample size and marker number was small, leading to an absence of signal, or visual inspection of the values of $L(K)$ would have identified runs of the MCMC with outlying values for $L(K)$.
- We further found the algorithm underlying the structure detects the upper most level of population, and that subgroups created by the best individual assignment produced by the structure permits the identification of sublevels of structuring [41].
- If the population structure and familial relatedness are not analyzed properly it may cause spurious associations (Table 3).

Wheat	Kernel weight	95 genotype	36 unlinked SSR	Without admixture. Allele frequency	Linear mixed-effects model (LME function)	F-test, at a level Alpha c	Default parameter	95th percentile of R^2 value	Alpha $c < 0.05$	Fst value estimation	STRUCTURE version 2.2	Bressanbello et al., (2006)
	Kernel area										TASSEL version 4.0	
	Kernel length										R	
	Kernel width										programmed	
	Superior milling										GENETIX	
	Score											
	Higher flour											
	Yield											
	Friability											
	Endosperm											
Separation index												
Wheat	Plant height	100 Winter varieties	5,525 DArT Marker including SNPs and PAVs	Principle coordinate analysis (PCoA)	LD, Best linear unbiased predictors (BLUP) genome association and prediction tool	False discovery rate (FDR) is calculated	MLM function	R^2 value is calculated	$p < 0.05$	Genome association and prediction tool is confirmed by GWAS	R package of R software	Bellucci et al. [14]
	Grain yield										TASSEL v 5.2.15	
	Bioethanol Production										TASSEL v 3.0.169	
Wheat	Tolerant to Pre-Harvest Sprouting	242 Genotype	250 SSR Markers	Without admixture and correlated allele frequencies	GLM and MLM	FDR is calculated	Structure based on Q matrix	R^2 value is calculated 0.56–4.48%	$p < 0.05$	No. of k is confirmed by posterior probability (ΔK)	STRUCTURE v 2.2	Jaiswal et al. [15]
	Moderately tolerant to PHS										MLM function based on $Q + k$ model	
	Susceptible to PHS											

(continued)

Maize	60 Agronomic traits including kernel	302 lines	89 SSR	K no. of fixed sub-population	GLM	Bonferroni correction	Q model, K model	R ² value is calculated 33–35 %	p < 0.01	Fst value estimation (p < 0.001)	STRUCTURE v 2.1 TASSEL GENETIX v 4.03	Flint-Garcia. [17]
	Protein		K no. of fixed sub-population	558 DArT 2,878 SNPs	GLM	Adjusted p-value	R ² value is calculated 2.3–3.9%	p = < 0.05	BOX-COS transformation	STRUCTURE v 2.1 TASSEL v 2.1	Roy et al. [6]	
	Starch											
	Oil											
	Moisture											
Spot blotch resistant	318 accessions	95 accessions	2,553 SNPs	Haplotype based method	Genomic control Bonferroni correction	GLM(Naive), Q, K and Q + K model is used	NA	p value is based on X ² test	PCA	STRUCTURE STRAT	Aramzana et al. [24]	
Plant	Traits	74 lines	290 SSR 30 random amplified polymorphic DNA (RAPD) 9 sequenced characterized amplified region (SCAR) (SCAR)									
Arabidopsis	Flowering time											
Vegetable	Pathogen											
	Resistant											
Pepper	Plant height											
	No. of fruit per Plant											
	Ten fruit weight											
	Total fruit weight											
	Fruit length											
	Fruit width											
	Pericarp thickness											

(continued)

Table 3 (continued)

Crops	Trait/s	Number of genotypes	Type and number of markers	Methodology population structure	Methodology association mapping	Multiple correction	Parameter ancestry model	R ² value	p value	Validation	Software used	References
Egg plant	Fruit weight	191 accessions	79 SNPs	NA	MLM (K + Q-model)	FDR is calculated	GLM (Naive-model), GLM (Q-model)	NA	$p < 0.001-0.05$	GWAS, cumulative density function is used for correcting the population structure	R package Tassel v4.0.25	Portis et al. [18]
	Fruit length											
	Fruit diameter (fdl1/4)											
	Fruit diameter (fdl1/2)											
	Fruit diameter (fdl3/4)											
	Fruit diameter (fdlmax)											
	Fruit diameter max											
	Possifion (fdlmax)											
	Fruit shape											
	Fruit curvature											
	Fruit apex shape											
	Peduncle length (cm)											
	Fruit calyx prickliness											
	Fruit calyx removal											
	Calyx coverage											
	Outer fruit firmness (Kg/cm ²)											
Inner fruit firmness (Kg/cm ²)												
Number of locules												
Flesh color												

Flesh green ring	
Plant growth habit	
Number of branches	
Leaf width (cm)	
Leaf length (cm)	
Adaxial leaf central	
Venation prickl.	
Adaxial leaf lateral	
Venation prickl.	
Abaxial leaf central	
Venation prickl.	
Abaxial leaf lateral	
Venation prickl.	
Stem prickliness	
Abaxial leaf prickles	
Number	
Adaxial leaf prickles	
Number	
Leaf hairiness	
Number of flowers/inflorescence	

(continued)

21 Conclusion

The population structure analysis defined the best groups of individuals within the group structure. However, ΔK emphasizes the correct number of clusters. Various genetic demands have gained a better hold, such as in choosing a better quality of individual for breeding programs and in the collection of germplasm bank accessions. Before starting AM, researchers should have knowledge of all genetic aspects of the germplasms and molecular markers. Through AM we can conduct genetic, physiological, and biochemical studies within individuals. The evolution of these genomic technologies continues to advance the debate of candidate gene versus genome. Originally, we had to search only a tiny fraction of the genome as needed. We expect to see more genome-wide association analysis and accept promising offers of complex trait dissection.

References

1. Rodney M (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nat Rev Genet* 2(5):370
2. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9(5):356
3. Álvarez MF, Mosquera T, Blair MW (2014) The use of association genetics approaches in plant breeding. *Plant Breed Rev* 38:17–68
4. Muluale T, Bekeko Z (2016) Advances in quantitative trait loci, mapping and importance of markers assisted selection in plant breeding research. *Int J Plant Breed Genet* 10:58–68
5. Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5(2):89
6. Roy JK, Smith KP, Muehlbauer GJ, Chao S, Close TJ, Steffenson BJ (2010) Association mapping of spot blotch resistance in wild barley. *Mol Breed* 26(2):243–256
7. Bressegello F, Finney PL, Gaines C, Andrews L, Tanaka J, Penner G, Sorrells ME (2005) Genetic loci related to kernel quality differences between a soft and a hard wheat cultivar. *Crop Sci* 45(5):1685–1695
8. Jannink JL, Bink MC, Jansen RC (2001) Using complex plant pedigrees to map valuable genes. *Trends Plant Sci* 6(8):337–342
9. Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol* 17(2):155–160
10. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38(2):203
11. Wen W, Mei H, Feng F, Yu S, Huang Z, Wu J, Chen L, Xu X, Luo L (2009) Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa* L.). *Theor Appl Genet* 119(3):459–470
12. Lu Q, Zhang M, Niu X, Wang S, Xu Q, Feng Y, Wang C, Deng H, Yuan X, Yu H, Wang Y (2015) Genetic variation and association mapping for 12 agronomic traits in indica rice. *BMC Genomics* 16(1):1067
13. Bressegello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172(2):1165–1177

14. Bellucci A, Torp AM, Bruun S, Magid J, Andersen SB, Rasmussen SK (2015) Association mapping in scandinavian winter wheat for yield, plant height, and traits important for second-generation bioethanol production. *Front Plant Sci* 6:1046
15. Jaiswal V, Mir RR, Mohan A, Balyan HS, Gupta PK (2012) Association mapping for pre-harvest sprouting tolerance in common wheat (*Triticum aestivum* L.) *Euphytica* 188 (1):89–102
16. Jaiswal V, Gahlaut V, Meher PK, Mir RR, Jaiswal JP, Rao AR, Balyan HS, Gupta PK (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(7):e0159343
17. Flint-Garcia SA, ThUILlet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* 44(6):1054–1064
18. Portis E, Cericola F, Barchi L, Toppino L, Acciarri N, Pulcini L, Sala T, Lanteri S, Rotino GL (2015) Association mapping for fruit, plant and leaf morphology traits in eggplant. *PLoS One* 10(8):e0135200
19. Riedelsheimer C, Lisek J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc Natl Acad Sci* 109 (23):8872–8877
20. Thomsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler IV ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 28(3):286
21. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904
22. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959. PMID: 10835412
23. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 3(1):e4
24. Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* 1(5):e60
25. Chhatre VE (2013) Population structure, association mapping of economic traits and landscape genomics of east Texas loblolly pine (*Pinus taeda* L.) Texas A&M University, College Station
26. Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 60(3):227–237
27. Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 68(2):466–477
28. Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MA, Hillel J, Mäki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* 159 (2):699–713
29. Beaumont M, Barratt EM, Gottelli D, Kitchener AC, Daniels MJ, Pritchard JK, Bruford MW (2001) Genetic diversity and introgression in the Scottish wildcat. *Mol Ecol* 10(2):319–336
30. Goossens B, Funk SM, Vidal C, Latour S, Jamart A, Ancrenaz M, Bruford MW (2002) Measuring genetic diversity in translocation programmes: principles and application to a chimpanzee release project. In: *Animal conservation forum*, vol 5(3). Cambridge University Press, Cambridge, pp. 225–236
31. Randi E, Lucchini V (2002) Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conserv Genet* 3(1):29–43

32. Arnaud JF, Viard F, Delescluse M, Cuguen J (2003) Evidence for gene flow via seed dispersal from crop to wild relatives in *Beta vulgaris* (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. *Proc R Soc Lond B Biol Sci* 270 (1524):1565–1571
33. Berry O, Tocher MD, Sarre SD (2004) Can assignment tests measure dispersal? *Mol Ecol* 13 (3):551–561
34. Cegelski CC, Waits LP, Anderson NJ (2003) Assessing population structure and gene flow in Montana wolverines (*Gulo Gulo*) using assignment-based approaches. *Mol Ecol* 12 (11):2907–2918
35. Caizergues A, Bernard-Laurent A, Brenot JF, Ellison L, Rasplus JY (2003) Population genetic structure of rock ptarmigan *Lagopus Mutus* in northern and Western Europe. *Mol Ecol* 12 (8):2267–2274
36. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. *Science* 298(5602):2381–2385
37. Karp A (1997) Molecular tools in plant genetic resources conservation: a guide to the technologies (No. 2). Bioversity International, Rome
38. Jarne P, Lagoda PJ (1996) Microsatellites, from molecules to populations and back. *Trends Ecol Evol* 11(10):424–429
39. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164(4):1567–1587
40. Pritchard JK, Wen X, Falush D (2010) Documentation for STRUCTURE software, version 2.3. University of Chicago, Chicago
41. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14(8):2611–2620
42. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2 (12):e190
43. Pritchard JK, Wen W, Falush D (2003) Documentation for structure software: version 2
44. Han S, Guthridge JM, Harley IT, Sestak AL, Kim-Howard X, Kaufman KM, Gilkeson GS (2008) Osteopontin and systemic lupus erythematosus association: a probable gene-gender interaction. *PLoS One* 3(3):e0001757
45. Tian C, Gregersen PK, Seldin MF (2008) Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet* 17(R2):R143–R150
46. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Hirschhorn JN (2005) Demonstrating stratification in a European American population. *Nat Genet* 37(8):868
47. Mir RR, Kumar N, Jaiswal V, Girdharwal N, Prasad M, Balyan HS, Gupta PK (2012) Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. *Mol Breed* 29(4):963–972
48. Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. *Trends Plant Sci* 10(12):621–630
49. Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol* 57(4):461–485
50. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 57(1):289–300
51. Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. *BMJ* 310 (6973):170
52. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature* 429:446–452
53. Garris AJ, McCouch SR, Kresovich S (2003) Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa L.*) *Genetics* 165(2):759–769
54. Simko I, Costanzo S, Haynes KG, Christ BJ, Jones RW (2004) Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. *Theor Appl Genet* 108(2):217–224

55. Zhang J, Zhao J, Xu Y, Liang J, Chang P, Yan F, & Zou Z (2015) Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor. *Front Plant Sci*, 6
56. Wei X, Jackson P A, McIntyre C L, Aitken K S, & Croft B (2006) Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. *Theor Appl Genet* 114(1):155–164

Genetic Mapping Populations for Conducting High-Resolution Trait Mapping in Plants



James Cockram and Ian Mackay

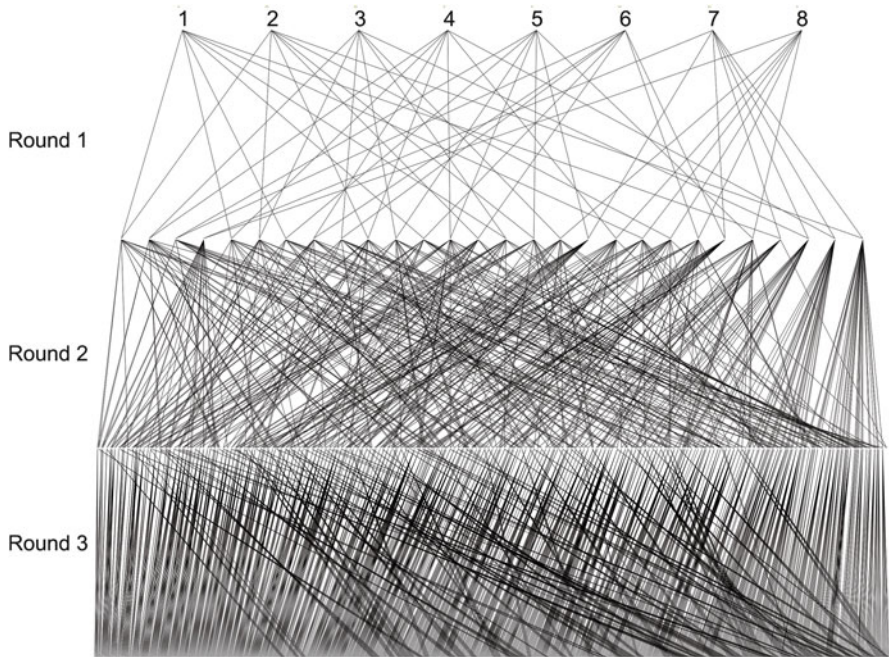
Abstract Fine mapping of quantitative trait loci (QTL) is the route to more detailed molecular characterization and functional studies of the relationship between polymorphism and trait variation. It is also of direct relevance to breeding since it makes QTL more easily integrated into marker-assisted breeding and into genomic selection. Fine mapping requires that marker-trait associations are tested in populations in which large numbers of recombinations have occurred. This can be achieved by increasing the size of mapping populations or by increasing the number of generations of crossing required to create the population. We review the factors affecting the precision and power of fine mapping experiments and describe some contemporary experimental approaches, focusing on the use of multi-parental or multi-founder populations such as the multi-parent advanced generation intercross (MAGIC) and nested association mapping (NAM). We favor approaches such as MAGIC since these focus explicitly on increasing the amount of recombination that occurs within the population. Whatever approaches are used, we believe the days of mapping QTL in small populations must come to an end. In our own work in MAGIC wheat populations, we started with a target of developing 1,000 lines per population: that number now looks to be on the low side.

J. Cockram (✉) and I. Mackay

The John Bingham Laboratory, National Institute of Agricultural Botany (NIAB), Cambridge, UK

e-mail: james.cockram@niab.com

Graphical Abstract



Keywords Arabidopsis multi-parent recombinant inbred line (AMPRIL), Fine-mapping, Genome wide association scans (GWAS), Linkage disequilibrium (LD), Multi-founder advanced generation inter cross (MAGIC), Nested association mapping (NAM), Power, Precision

Contents

1	Introduction	111
2	The Importance of Mapping Trait Loci	111
3	The Importance of Fine Mapping	112
4	Factors Determining Precision in Fine Mapping	113
4.1	Recombination	113
4.2	Population Size	114
4.3	Size of Effect	114
5	The Need for Replication	115
6	Association Mapping	115
7	Genome-Wide Association Studies (GWAS) Compared to Experimental Populations for Fine Mapping	116
8	Experimental Populations	118
9	Biparental Populations	119
10	Bulk Segregant Analysis (BSA)	120
11	Advanced Intercross (AIC)	121

12	Near Isogenic Lines (NILs)	122
13	Multi-Founder Populations	123
	13.1 Nested-Association Mapping	124
	13.2 Heterogeneous Stock	125
	13.3 MAGIC	126
14	Analysis Approaches for MAGIC	129
	14.1 Arabidopsis Multi-Parent Recombinant Inbred Line (AMPRIL)	131
	14.2 Linked or Multiple Mapping Populations	132
15	Conclusion and Outlook	132
	References	134

1 Introduction

In this chapter, we review the use of mapping populations for precision location of quantitative trait loci (QTL). We focus on experimental populations created explicitly for trait mapping, distinguishing these from collections of lines or individuals used in the association-mapping approaches described in Chapter 5. We start by discussing trait mapping and the need for fine mapping, next describe some general properties and requirements of approaches to fine mapping, after which we describe specific approaches. We finish by considering prospects for the future.

2 The Importance of Mapping Trait Loci

The tagging of QTL with genetic markers has a long history [1]. However, most progress has been made since the 1980s with the development of DNA markers initiating the era of genome scans in bi-parental crosses [2, 3]. This has continued to the present, as new, cheaper classes of genetic markers (e.g., [4]) and improved statistical methods and software (e.g., [5]) have become available. Markers tagging QTL can be used to ease introgression of novel variation from un-adapted germplasm into elite lines, for marker-assisted selection, and for stacking multiple sources of disease resistance [6]. However, trait mapping has also been tempered by cautious voices, alerting people to the risks of bias in estimating effects [7], of lack of precision of QTL location [8], and of relevance to plant breeders' germplasm. Nearly 10 years ago Bernardo [9] commented on the very poor transfer rate of results from trait-mapping experiments to breeding programs. New population and genomics resources and more widespread understanding of the requirement for statistical power are now focusing effort on studies that map more QTL to smaller intervals in breeder-relevant germplasm. Our hope is that this will result in better application of QTL within breeding programs and other studies in the future, rather than being an end in itself.

3 The Importance of Fine Mapping

Fine mapping is the process by which the location of a QTL is reduced from an initial interval of 20 cM or more to an interval of a few cM or less. Taken to its final conclusion, fine mapping can lead to the identification of the causative genomic lesion, which could be a single nucleotide polymorphism, SNP, or other variant. Fine mapping has merit in biological studies: clearly if the causative polymorphism is identified, then the door is opened to further detailed molecular characterization and functional studies of the relationship between polymorphism and trait, and the wider genetic network. However, reduction of a QTL linkage interval to a smaller region without exactly identifying the functional polymorphism or gene is also of value. Near isogenic lines (NILs) can be created containing the alternative states at the QTL locus and used for detailed phenotyping. NILs cannot discriminate between pleiotropy and close linkage but the smaller the interval, the lower the chance of misinterpretation. For example, fine mapping in rice showed a rice photoperiod sensitivity locus, previously known as *Heading date 3* (*Hd3*), could be genetically dissected into two closely linked loci *Hd3a* and *Hd3b*, allowing allele-specific effects at both loci to be discriminated in NILs [10]. Fine mapping is also of direct relevance to breeding since it makes QTL more easily integrated into marker-assisted selection programs. A QTL located to an interval of 20 cM requires that a 20 cM tract of chromosome is tagged during selection and backcrossing. This prevents recombination within the region and the potential creation of favorable haplotypes containing the QTL and other, unidentified loci. In addition, wide intervals reduce the potential to stack favorable QTL since even with a small number of QTL, it is quite likely that intervals will overlap.

Genomic selection [11, 12] is the process whereby statistical models are applied to large numbers of genetic markers to predict the breeding (or trait) values of candidates for selection which are not themselves phenotyped. Because selection and phenotyping are decoupled, very high intensities of selection are possible (because single individuals rather than cultivars are selected) and reductions in cycle time can be made (because the breeding cycle is from parent to progeny, without time taken for cultivar development). In the near future, genomic selection will become routine in plant breeding [13]. QTL, known or recently discovered, can be incorporated into genomic prediction equations in an optimum manner, but a QTL characterized only by loosely linked flanking markers eliminates large tracts of chromosome from inclusion in trait prediction algorithms with potential loss of precision.

Identification of functional polymorphisms also allows genome-editing technologies, such as clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated protein 9 (Cas9) [14], to introduce favorable changes directly into plant genomes without any subsequent linkage drag. Methods to increase the number of functional polymorphisms identified could have a substantial influence on rates of genetic improvement [15]. Breeding programs have already been studied in which genome editing is used to this end [15].

4 Factors Determining Precision in Fine Mapping

The precision with which a QTL can be located depends on three related factors: the recombination fraction between it and the available genetic markers, its heritability, and the size of the mapping population. These factors are general across all types of mapping populations. Other factors can affect specific approaches to fine mapping and will be discussed in context as they arise.

4.1 Recombination

Trait mapping in plants exploits the correlation between genetic markers and QTL. A zero correlation is expected between unlinked markers, rising to an expected maximum for a marker co-located with the QTL. However, the observed maximum is very unlikely to be located precisely at the QTL position. For example, for a marker and QTL separated by a recombination fraction of 0.01, the probability of observing at least one recombination in 100 meioses is only 63%; there is a very high chance that none occur, in which case a marker located at the QTL and a marker located ~1 cM away cannot be distinguished. In practice, the discrimination is likely to be much worse. With 1,000 meioses, the chance of observing no recombinations with a recombination fraction of 0.01, is 0.99996. Populations for fine mapping therefore require high levels of recombination. This is achieved in two ways. Firstly, population size can be increased, and secondly, the mapping population can be created over multiple generations of crossing. The early generations of selfing to generate recombinant inbred lines, when the frequency of double heterozygotes is reasonably high, also provide some opportunity for recombination. For this reason, fully inbred lines give greater precision than doubled haploid (DH) lines.

There is a conflict between power and precision in fine mapping. Power to detect a QTL is increased in populations with reduced recombination. This is exploited in standard bi-parental populations, but also, for example, in the pre-QTL mapping era by methods such as use of anisoploid lines (which possess a chromosome number that is an odd multiple of the haploid number, e.g., triploid) and whole chromosome substitution lines (e.g., [16]) which allowed the location of major effects (or the cumulative effect of many small effects) to whole chromosomes or chromosome arms. To increase precision, the mapping population requires more recombination events, but this comes at the expense of power. As a simple example: suppose a genome-wide significance threshold of 0.001 is required to detect a QTL linked to a marker located at a distance of 10 cM in an F_2 , and that additional cycles of crossing were used to increase precision to 1 cM. Ignoring complications from interval mapping and other multi-marker approaches, this would require a tenfold increase in marker density and therefore a tenfold change in the genome-wide Bonferroni corrected p -value to 0.0001. To maintain power to detect such an effect, population

size would need to be increased (see below). Adopting a lower significance threshold is an alternative, and this may be possible if the study population is not being used for GWAS, but to fine map a QTL identified in other studies.

4.2 Population Size

Population size has two effects. Firstly, as described above, bigger populations capture more recombination and therefore offer greater precision. However, in addition, bigger populations have greater power to detect QTL. For example, in a mapping population in which QTL and markers are segregating at a frequency of 0.5, for example a DH population or population of recombinant inbred lines (RILs), the power to detect a QTL accounting for 10% of the phenotypic variation between lines at a significance threshold of $p \leq 0.001$, and assuming a perfect marker for the QTL, is 0.52 with a population size of 100. Increasing the population size to 200 increases the power to 0.92. Increasing population size, therefore, increases both the precision with which phenotypic effects of marker classes are estimated and also increases the power to detect QTL. It is for this reason that increasing population size is preferred over increasing replicate number for a fixed population size. This latter approach increases the power with which QTL are detected but has less effect on precision.

4.3 Size of Effect

Bigger effects are easier to detect and are more precisely located. In the extreme, a QTL may be so large that, in practice, it behaves as a completely penetrant major gene: essentially as a marker itself. At the other extreme, for a highly polygenic trait, the effect of any individual locus may be very small and there will be very little power to detect or locate QTL. This is true even if the heritability of the trait is very high: it is the heritability of the QTL effect itself that is the dominant determinant of power and precision. However, for traits of low heritability, replication can be used to increase the heritability of line means and thus increase power and precision, though as previously mentioned precision is better increased by increasing recombination. There is a counter view, however [17], that conventional QTL mapping is very successful in locating QTL precisely, though the examples given are what most workers would regard as large QTL effects.

5 The Need for Replication

Whatever the interval reported or the method used, there is a requirement for replication, or validation, of the observed effect prior to progression of the QTL into marker-assisted breeding or functional investigation. Such studies should preferably be in a completely different population to that in which the QTL were first identified: it is not appropriate, for example, to partition a collection of germplasm, or a mapping population, and use one portion for “discovery” and the other for “validation.” Such an approach amounts to cross validation, and may reduce the rate of false positives *within that study* or result in reduced bias in the estimation of genuine effects through elimination of the winner’s curse [18], but it is not an independent study. The winner’s curse is the phenomenon that evidence of a new effect, provided by significance testing, often gives an inflated estimate of the size of that effect. In linkage analysis, this is often referred to as the Beavis effect [7]. Replication in independent studies has become a standard for publication of association-mapping studies in some journals [19] and has been more widely advocated, e.g., [20, 21]. In wheat, for example, we have identified 26 QTL associated with yellow rust resistance in a large panel of 488 lines, of which 11 out of 13 of the strongest associations were replicated in independent bi-parental mapping populations created for that purpose. Out of nine hits in the original association-mapping panel that were statistically significant at a less stringent significant threshold ($p \leq 0.001$), only three were replicated. Breeders were substantially more confident of incorporating the replicated QTL into their marker-assisted selection programs. Distrust of results from single studies contributes to the relative lack of uptake of QTL into marker-assisted selection programs [9].

6 Association Mapping

Association mapping, also known as linkage disequilibrium mapping, detects and locates QTL based on the strength of the association (correlation) between genetic markers and the traits under study. It relies on the magnitude of linkage disequilibrium (effectively the correlation) between genetic markers and QTL declining rapidly with genetic distance. Detection of a strong correlation between a trait and a genetic marker is therefore taken as evidence that a QTL is in close proximity to the marker. Association mapping can, in principle, be applied to any population or collection of lines or individuals, and in general will give higher precision than found in bi-parental populations. The pattern of LD decay is remarkably similar in diverse populations, due to the mechanism by which LD decays over time: at a rate of ($one - the\ recombination\ fraction$) per generation. However, differences in scale will be seen, due to differences between populations and species in the forces that

create LD – mutation, selection, and drift – and the age of the population or collection of lines under study with reference to their shared genealogy.

However, virtually all populations or panels assembled for association mapping include some degree of population structure or subdivision arising from ancient and very recent differences in the shared ancestry of the lines in the collection. If these are not taken into account, very high frequencies of false-positive results can arise: false in the sense that the observed association, though genuine, has arisen from some other cause than the close linkage of marker and QTL. Fortunately, statistical methods, in particular the use of the mixed model, can robustly adjust for population structure effects. For example, in the *Triticeae*Genome association-mapping panel of European wheat [22], 11% of squared correlations between markers pairs ≥ 0.8 were among unlinked markers. However, with simulated traits, application of the mixed model gave good control of false positives but identified that a higher marker density was required to improve precision. Power calculations, and associated estimates of expected precision, should always be reported in association-mapping studies. Association-mapping studies have therefore become routine in plants. For major genes, accuracy can be to the gene level (e.g., [23]), though independent evidence of the functionality of the candidate should also be obtained, for example through transformation or reverse genetics studies. For QTL that do not account for most of the genetic variation, accuracy is lower, but is generally greater than seen from bi-parental mapping populations. However, association mapping is not a panacea and other approaches still have a role, and may be better under some circumstances.

7 Genome-Wide Association Studies (GWAS) Compared to Experimental Populations for Fine Mapping

Genome-wide association studies (GWAS) have become increasingly popular in plants: they are easy to set up, requiring only the collection of pre-existing lines or cultivars with no need for de-novo crossing and selfing [24]. They also often come with pre-existing phenotypic data [23, 25]. However, there are a few notes of caution to be made:

GWAS studies are often proposed because of their assumed increased precision for locating QTL. However, this is often not realized. A demonstration that linkage disequilibrium (LD) decays quickly is usually used as an indication of likely precision in mapping, but if QTL are to be discovered in a genome scan, very large population sizes are required for fine mapping. Often, the sizes used in published studies are ludicrously small and there is no accompanying estimate of power. Occasionally, both population size and marker numbers are risibly low; 46 SSRs and 30 accessions with accompanying claims of high power and precision is the worst example we have seen published in an otherwise reputable journal, in this case with no report of the rate of LD decay or of power. Low power to detect a

QTL must throw doubt on published statements about the frequently large number of QTL detected [26]. This problem is not overcome by establishing, as often reported, that putative hits lie in known QTL linkage regions. Given the long history of QTL mapping, it is difficult for a locus not to lie in any region of the genome chosen at random. Statements, therefore, that say some proportion of the detected marker trait associations are new and the rest replicate previous results require greater statistical support, which can be easily calculated or simulated, but this is seldom done in plants. An excellent example in *Drosophila* is given by Highfill et al. [27]. They identified five QTL for variation in lifespan in a multi-parent advanced generation inter-cross (MAGIC) population. Over 100 QTL for this trait had previously been identified and they established through simulation that the probability that *all* of five randomly located QTL would overlap with one or more of these 100 was 0.85.

The problem of low power may be insurmountable. Population sizes can be unredeemably small, especially in the public sector where the only germplasm available may be from collections of varieties released commercially by breeders. If the number available is too low, the only option may be to create more, in which case an experimental population such as MAGIC may be a better alternative.

The statistical control of population structure and kinship works well in controlling the false-positive (Type I error) rate, but this can be at the expense of power. If a QTL is highly associated with a major population subdivision (e.g. [28]), it may be difficult to detect, or in extreme cases, not be detected at all: adjustments for kinship and population structure will reduce the power to detect any association between a QTL and linked marker that is also correlated with those effects. This can be a major problem: an attempt to increase power by capturing lines grown over a greater geographical and temporal range inevitably also increases population structure within the dataset. For example, the *Triticeae* Genome panel [22] includes lines of British, German and French origin. German lines tend to be taller: height being controlled by the use of growth regulators rather than through semi-dwarfing genes, and French lines tend to be earlier flowering – to avoid summer drought stress. Therefore, the frequency of major flowering time and height-reducing loci differs between countries so power to detect these is also reduced. In this case, these loci are of such great effect that they were still detected, though it is quite possible that minor QTL affecting these traits were not.

In addition, within narrow temporal and geographical ranges, although the problems of population structure may be reduced, LD may decay quite slowly with genetic distance; a consequence of close kinship among all lines. As a result, the precision with which QTL are detected can be reduced.

Power to detect and locate QTL in association-mapping panels also depends on the allele frequency of the QTL. Very rare alleles will not be detected in genome scans, even if the function polymorphism itself is tested and the effect is large. The detection of rare variants is an acknowledged problem in human genetics [20]. In plant science, we have the alternative of making experimental crosses and populations to reduce this problem, though success depends on the identification of appropriate founders.

Many of the issues surrounding the use of association-mapping panels for fine mapping can be avoided if the stringency of statistical significance is reduced. This is justified if the purpose of the experiment is not a GWAS but rather to test a small number of candidate genes or polymorphisms, or to fine map a genomic region for QTL identified in a previous study. For example, a genome scan with 10,000 markers would require a Bonferroni-adjusted significance threshold for an experiment-wide p -value of 5% of 0.0005%. Testing 100 candidate polymorphisms would require 0.05%.

In spite of the problems, association mapping is a powerful tool for fine mapping and notable successes have been reported [21, 23]. There is still of course a role for other methods: in particular using bespoke experimental populations.

8 Experimental Populations

Experimental populations for trait mapping pre-date the use of association-mapping panels, though, in essence, the principles are identical: QTL are located by the strength of the association between markers and traits, exploiting LD. In most experimental populations, LD decays slowly with genetic distance: a consequence of the limited number of generations between the creating of extensive LD by crossing a small number of founder or parental lines, and the small number of generations from that event before mapping takes place.

Experimental populations circumvent many of the problems with the use of association-mapping panels, but introduce problems of their own. There is little or no effect of population structure within them, so the correct Type I error rate is usually achieved. They tend to have higher power to detect QTL than association-mapping panels – a consequence of the slower rate of decay of LD with genetic distance. Correlated with this, they can lack precision in locating QTL in comparison to similar size association-mapping panels. In addition, they generally lack genetic diversity compared to association-mapping panels.

Selection of parents or founders is of great importance in experimental populations. This will be discussed in the context of individual population types below. As general principles, however, we note that parents or founders can be selected for similarity of phenological traits, in the hope of reducing segregational variation in the population for those traits to ease the phenotyping of the traits of major interest. For example, in many cereal species, lodging and resistance to abiotic stresses such as drought are affected by the developmental phase of the plants. Reducing segregational variation in, for example, flowering time, may increase both the power and precision with which QTL for stress are identified – in essence by increasing the heritability of the QTL for stress *per se*. However, matching parents on phenology is absolutely no guarantee that the problem is eliminated: dispersion of alleles with variable effects on phenology among the parents will generate substantial, and commonly transgressive, segregational variation in the progeny. In addition, selection of parents in this manner may eliminate

variation at linked loci, particularly if, as seems inevitable, there are interactions between phenology and other traits. The opposing view is to adjust for phenological variation during analysis by inclusion of covariates, or through partitioning the population into subsets with matched phenology (“slicing and dicing”), and then carrying out QTL analysis only within subpopulations. “Slicing and dicing” is equivalent to inclusion of covariates in the analysis to account for subgroup membership: in essence it amounts to binning a quantitative phenological factor such as flowering time into covariates with values of 0 and 1, and is therefore crudely equivalent to simply including phenology traits directly as covariates. Our opinion is that the inclusion of covariates is the better approach and is more flexible, plus it is likely that this will be required with matched parents anyway. However, as far as we are aware, the approaches have not been compared empirically or by simulation. In our own work, we have found the presence of transgressive segregation to be useful in interpretation rather than a hindrance in phenotyping.

9 Biparental Populations

Mapping in the progeny or lines from a cross between two inbred lines remains the standard approach for genetic mapping in plants (Chapter 4). Mapping usually takes place among DH lines derived from the F_1 , or RILs derived from the F_2 or backcross generations. Mapping directly among F_2 individuals is rarely used, partly because the genotypes of individuals are often poorly assessed by their phenotype (due to low heritability) and partly because, in the absence of clonal propagation, the individuals cannot be maintained indefinitely for annual or biennial species. The strength of the standard bi-parental population is in its power: LD decays slowly within chromosomes and there is no expectation of LD between loci on different chromosomes (LD between chromosomes is eliminated by the non-random mating of the parents: only the cross is made, not the selfs. With random mating, substantial LD would be found between chromosomes). As ever, the higher power is associated with low precision: Kearsey and Farquhar [8] state that precision for a standard QTL is to a region of 10–30 cM. They also make the point that the addition of markers to bi-parental mapping populations beyond a density of about one per 15 cM has limited effect on precision. The same point was made by Darvasi and Soller [29], who state that no increase in precision is made once an inter-marker recombination fraction of 0.1–15 is achieved. However, in practice more markers are commonly used: although only modest numbers of evenly spaced markers are required, it generally takes a much higher density of markers in the same population to produce an accurate genetic map in the first place. Moreover, reduced marker costs, SNP chips, and cheap genotyping by sequencing (GbS) will make discussions of marker density of historical interest. However, Kearsey and Farquhar [8] and Darvasi and Soller [29] point that it is more meiosis that is required to increase precision remains valid: small mapping populations, even if they have adequate power to detect QTL, lack in precision for QTL of modest size.

Selection of parents to create bi-parental mapping populations is generally trait driven: they are often selected as contrasting extremes for the trait of interest. This is a strength, in so far as it increases the likelihood that the population is segregating for multiple QTL for the trait. It also increases the chance of capturing alleles that are rare in the population sampled (and so are unlikely to be detected in GWAS studies). However, it is also a weakness in that the favorable alleles mapped may already have been fixed in breeders' germplasm. The trait-focused nature of many bi-parental mapping populations also means they tend to have short-term use. Interest in mapping additional traits is better served by creating additional targeted crosses, though there are exceptions of populations that have been more widely used. For example, the wheat Avalon \times Cadenza mapping population has been used a lot to map many traits [30]. Nevertheless, if inbred parents are sampled at random from a population, at most only half the loci segregating in the population will be segregating in the cross – assuming two alleles at equal frequencies at all loci – and it could be a much lower proportion.

One method of reducing the cost of creating bi-parental mapping populations is to use Rapid Bulk Inbreeding (RABID) [31, 32]. Here, inbreeding takes place in bulk, so is cheap, with modest numbers of markers (~ 100) used after inbreeding to identify a set of individuals with minimal relationships as the mapping population. As bulk inbreeding is cheap, it is possible to create multiple populations speculatively, but only genotype them, and multiply selected lines for phenotyping as required. A similar use of markers to select the most unrelated set of individuals could also be applied to lines produced by single seed descent or DHs, also providing the opportunity, for example, to eliminate lines carrying un-recombined chromosomes. Such chromosomes increase power but clearly do nothing for precision! We are not aware of any research looking at this.

10 Bulk Segregant Analysis (BSA)

Bulk segregant analysis (BSA) has had a revival now that DNA and RNA resequencing is routine in most plant species (reviewed for crops by Zou et al. [33]). The principle is that contrasting extremes are selected from a population and genotyped in bulks. Markers that distinguish between the bulks are judged to be closely linked to causative polymorphism. Bulking of the extremes saves money on genotyping, but is not necessary: genotyping individuals rather than bulks is more accurate.

There are three general approaches in terms of the biological materials used: (1) F_1 -derived individuals (F_2 s, RILs, DHs), e.g., identification of metabolite QTL for flavonoid production in *Arabidopsis* [34], (2) EMS-induced mutations, such as targeting induced local lesions in genomes (TILLING) populations, and (3) pooling genetically diverse breeding lines, e.g., red versus green hypocotyl in sugar beet [35].

Although the underlying principle of BSA is very simple, the expected composition of the selected pools, taking into account population size, number of selected individuals, genotype frequencies in the population, and genetic and environmental effects, is surprisingly complex [36]. Mackay and Caligari [37] used the methods of Hill [36] to compare BSA in F_2 and backcross generations, concluding that the F_2 was to be preferred if screening with dominant markers. There is, perhaps, a need to study strategies and population types more extensively, particularly in the light of DNA resequencing for genotyping: it is telling that the Hill [36] paper has only been cited four times.

11 Advanced Intercross (AIC)

The first experimental design intended specifically to improve precision in mapping was the Advanced Intercross (AIC) of Darvasi and Soller [29]. Acknowledging that absence of precision came from the limited recombinations captured in most bi-parental populations, they proposed increasing this by intermating the F_2 for several generations prior to mapping. For example, they stated that eight additional cycles of random mating could reduce the confidence interval of a QTL from 20 to 3.7 cM. Recent examples of AIC in plants include genetic investigation of female control of non-random mating in *Arabidopsis* (four rounds of intercrossing, 490 RILs, [38]) and disease resistance in maize (four rounds of intercrossing, 302 RILs [39, 40]). AIC, however, is used surprisingly rarely for crops. One reason for this may be the time required to make the additional crosses – especially if the mapping population was established in the first place only to detect (and locate) QTL for a specific trait in a short-term project, as is often the case. Another reason is that whereas a DH-mapping population may be created reasonably economically from an F_1 , since multiple DH lines can be created from a single individual, to do the same with an advanced intercross, or just an F_2 , requires, ideally, that a single DH is produced from as many outcrossed individuals as possible. Using the technologies currently available in many crops, this is not practical. Note that although the AIC will increase precision, this is at the expense of power. Essentially, this is a multiple testing problem. The number of independent genomic intervals, or their effective equivalent, which must be tested for the presence of a QTL in an AIC, is greater than in a simple cross. The genome-wide significance threshold must therefore rise. As an approximation, if QTL are located to 20-cM intervals in an F_2 but to 4-cM intervals in an AIC, then a genome-wide significance threshold of $-\log_{10}(p) = 3$ (say) would need to be increased to 3.7 in the advanced intercross. It is possible that a two-stage strategy could be derived in which QTL detection is first undertaken in a standard population with a lenient significance threshold with validation and fine mapping of these intervals occurring in an AIC as a second stage. The optimization of such a mapping strategy across two populations, which could also involve selective genotyping and phenotyping, merits study.

12 Near Isogenic Lines (NILs)

Near isogenic lines (NILs) are derived via repeated backcrossing of genetically distinct parental lines, most commonly with the aim of transferring a single chromosomal region of the donor parent into the genetic background of the recipient parent (reviewed by Kooke et al. [41]). Ultimately, over several generations of backcrossing, selection, and phenotyping, the interval containing the QTL should be eroded to something quite small. In practice, backcrossing beyond the second or third backcross is seldom carried out, the interval containing the QTL can remain large, and significant additional regions of donor parent genome can remain in the genetic background. The main use of NILs is that the effect of the QTL (and surrounding genome) can be characterized in much greater detail simply by phenotyping two lines rather than the whole, or a substantial part, of the mapping population. The power of NILs to detect a phenotypic effect of the QTL may not be substantially greater than from a similar size: if the QTL is the sole cause of genetic variation in the cross, power will be identical for identically-sized experiments. However, for a trait with a heritability of 50% of which the QTL accounts for 10%, the size of the mapping population would need to be 1.67 times greater than the experiment with NILs to get the same power.

Additionally, NIL pairs can be crossed, for example to fine-map a single QTL within the interval, or to combine two or more regions to investigate combined QTL effects. Collections of NILs that collectively capture all of the donor parent genome can be used for genetic mapping. Generation of such NIL populations, also termed chromosome segment substitution lines (CSSLs), can be aided via the use of molecular markers. Here, markers are used to both select for specific donor regions (foreground selection) and select against unwanted donor regions elsewhere in the genetic background (background selection), ultimately leading to an inbred population that can be used for genetic mapping. A recent example is the development of NIL populations in *Arabidopsis* (75 NILs [42]).

The backcrossing and purification required to develop NILs take time, leading to lag in their creation following QTL discovery by some other means. In mapping populations of inbred lines produced by selfing, inbreeding is seldom complete; the probability that an inbred line is fully homozygous after six generations of selfing is 0.03 for a species with 21 chromosome pairs and a total map length of 17 Morgans [43]. At this stage, the average line will contain 3.5 heterozygous tracts of chromosome and the total length of the genome that is heterozygous will be 27 cM on average. As a result, it is likely that individuals who are heterozygous, or families that are still segregating, for any tract can be found. Such heterogeneous inbred families (HIFs) can be used to rapidly create NILs through selfing [44]. Essentially the same approach was been used by Yamanaka et al. [45], for example, to fine map the *FTI* locus for soybean flowering time to a distance of 0.1 cM from the closest marker. In this case, a fully homozygous F_8 inbred line, aside from a 17 cM heterozygous tract around the *FTI* QTL was identified and 18 F_9 individuals

from the same line were selfed to create a population of $>1,006 F_{10}$ individuals that were used for mapping.

The utility of HIFs for power and precision is, of necessity, restricted to mapping populations produced by selfing rather than through DH. The probability of detecting HIFs for any particular QTL will depend on the generation of selfing. In the case of Yamanaka et al. [45], assuming the F_8 family originated from a single F_7 individual, there is a 0.0156 probability that the F_8 is segregating. Yamanaka tested 210 plants so the probability of *not* finding a segregating family is $(1-0.0156)^{210}$ or 3.7%: even with deep inbreeding, provided population sizes are modest, it is likely that HIFs will be found in bi-parental mapping populations and can be used to create NILs. This opens up the possibility of creating a tiled array of HIF's covering the genome.

13 Multi-Founder Populations

Within the last few years, and in parallel with the advent of association mapping for crops, more complex mapping populations have been advocated and used. These benefit from capturing increased levels of genetic diversity from the use of multiple founders. The use of multiple founders can also allow the incorporation of LD relationships among the founder into the analysis, with a potential gain in precision. These properties make multi-founder populations well suited to investigation of multiple traits, and are therefore used as community resources for genetic research. It is a feature of multi-founder populations that they are large: as a second-generation approach to mapping, it has been recognized that small populations are underpowered: a problem that is compounded in fine mapping where experimental methods to increase precision generally reduce power of QTL detection.

Multi-founder populations differ conceptually from bi-parental populations. Greater effort is required in their creation and they are generally designed to map multiple QTL for multiple traits. They are not a "use once and throw away" resource in the same way that bi-parental mapping populations can be. Choice of founders is important. The expectation is that inferences and discoveries in a multi-founder population will be applied to a wider population. An explicit understanding of the extent or range of this population is likely to give rise to the best choice of founders. For instance, selecting a set of lines to maximize global diversity is not the best strategy if the prime interest is to map traits for productivity in one particular agro-ecological environment. As with bi-parental mapping populations, there would be a risk, though reduced, of mapping QTL for loci for which the favorable allele is already fixed in the target environment. However, if the prime interest is in positional cloning, the best strategy may well be to select a set of founders that maximize species or crop diversity, though conditioned by the understanding that the lines that are produced will require phenotyping. This could cause problems in crosses between wild and cultivated forms, for example, or at least limit the range of traits that can be scored.

13.1 *Nested-Association Mapping*

Nested-association-mapping (NAM) populations, first proposed by Yu et al. [46] for the outcrossing species maize (*Zea mays*), are based on crossing multiple inbred lines to a single reference line then deriving multiple bi-parental sub-populations, either as DH lines or RILs. Originally designed for outcrossing species where LD decay is rapid (in maize, LD decays to background levels within ~1 kb), NAM combines the benefits of linkage mapping and association mapping: the linkage analysis within populations gives power to detect loci without the need for very high marker coverage, while the exploitation of more rapidly decaying LD across the multiple founders improves precision [46]. Protection against false positives arising from population structure is provided, in effect, by testing for association in the presence of linkage, an approach analogous to that of the QTDT [47, 48] in human genetics. NAM can capture high genetic variation while avoiding the complications of population structure, as usually found in GWAS panels.

The first NAM population was published in 2009, consisting of 25 maize inbred lines crossed to a single recurrent parent, resulting in 200 RILs per sub-population, and 5,000 RILs in total [49]. Since then, the maize NAM population has been extensively used for genetic analysis of multiple traits, including morphological, disease resistance, and metabolite phenotypes (e.g., [50–58]). Subsequently, NAM populations have been created in sorghum [59], as well as the inbreeding species, barley (*Hordeum vulgare*) [60] and wheat [61, 62]. While the classic NAM population design as exemplified by McMullen et al. [49] has invariably been used to date, related designs have been advocated. Simulation using empirical data in maize and *Arabidopsis* has shown that given a fixed total population size, power and reduction of false positives is optimized via employing designs such as diallel or factorial designs for outbred species [63].

For inbreeding species, where consideration was given to the effort needed to achieve the crosses required, double round robin design was found to be a good alternative [63]. Recently, an advanced backcross NAM (AB-NAM) population has been developed in barley that introgresses wild barley landraces into the exotic background [64]. The populations were developed by backcrossing 25 wild barley accessions to an elite barley cultivar. The lower proportion of wild genome in the recombinant lines makes phenotyping and mapping of loci in unadapted material easier.

NAM populations involve fewer crosses than alternative designs such as MAGIC (discussed below), and additional crosses can be added over time. Moreover, NAMs can emerge as a bi-product of other breeding and research activities. For example, the Wheat Improvement Strategic Programme (WISP) [65] is creating novel allohexaploid germplasm in wheat (synthetic wheat) by crossing tetraploid wheat with wild diploid goat grass (*Aegilops tauchii*). New synthetics are backcrossed to two elite lines and recombinant inbreds produced. This work was initiated purely for pre-breeding, but in essence has also created a NAM similar to that of Nice et al. [64]. A similar exercise in the WISP is producing lines from

backcrosses of landraces to elite wheat; also producing a NAM. Care must be taken with these emergent resources, however, to be aware of their statistical power. As they are not necessarily designed for mapping in the same way as the maize NAM, the numbers of lines created may be quite low. Power in mapping [66], as in life [67], is always worth thinking about.

Although multiple parental lines are involved in NAM, the creation of haplotype diversity is limited. With 26 lines at most 50 recombinant haplotypes can be created between two loci (out of 67,108,864 possible, assuming the parental loci are all difference). A greater limitation may be that these 50 will always involve the common parent: no novel haplotypes between the 25 unique parents are generated. Guo et al. [48] proposed that NAMs were created with two recurrent parents to avoid the emphasis on detection of QTL for which the recurrent and non-recurrent parents differ. This would create more haplotype diversity (100 recombinant haplotypes), but the potential for generating novel haplotype diversity is not as great as with MAGIC populations, as discussed below.

Where founders have been sequenced, lower-cost genotyping approaches (such as SNP arrays, GbS, or low-pass sequencing) in the progeny allows founder genotype to be projected onto the progeny. The prospect of sequencing the genomes of all individuals to medium-to-high coverage within a mapping population is now being realized. To date this has occurred in diversity and association-mapping panels, as this approach provides a wide survey of genetic diversity within a species. Examples include 3,000 rice accessions to $14\times$ sequencing depth [68] and 80 accessions to $\sim 15\times$ depth within the *Arabidopsis* 1,001 Genomes Project [69]. The continued development of sequencing technologies means that the application of genome re-sequencing to other types of plant mapping populations is inevitable in the near future.

13.2 *Heterogeneous Stock*

Mott et al. [70] proposed the use of an outbred mouse population (heterogeneous stock), created from eight inbred laboratory strains intercrossed for 60 generations, for fine mapping. Subsequently, Valdar et al. [71] mapped 843 QTL for multiple complex traits including aspects of behavior, which were located to 95% confidence intervals of ~ 2.8 Mb on average. Historically, development of heterogeneous stock in the mouse was instrumental in initiating the development of the first MAGIC populations in *Arabidopsis* and wheat. However, the mouse heterogeneous stock was not developed initially for mapping. Similar populations exist in crops, also not developed for trait mapping. In outbreeding species, including many crops, populations are often maintained in isolation for recurrent selection programs. For highly heritable traits, these can be used directly for mapping, though for traits with low heritability, clones (or inbred lines) would need to be extracted for phenotyping first. Similar populations exist in inbreeding species too. In wheat, for example, a French population was established with 60 founders, segregating for

genetic male sterility [72]. Provided seed is only ever harvested from the male sterile individuals, the population is maintained in a crossbred state. Population structure effects should be minimal, though there will be close familial relationships, so the risk of spurious marker trait associations should not be high. 1,000 inbred lines have been extracted from the French population after 12 generations of outcrossing. The population had been maintained as a bulk plot grown outside, so natural selection occurred, following the principles of Dynamic Management [72]. Thépot et al. [73] reported the first results from the French population, detecting 26 genomic regions under selection, of which six were associated with flowering time.

13.3 *MAGIC*

Following the success of the heterogeneous stock in mouse, we advocated the development of similar resources in crops – which we termed MAGIC populations [25]. These are in essence identical in construction to the mouse Collaborative Cross [74], with the exception that inbred lines are typically produced by selfing rather than by sib-mating (all that is possible in animals). Since then, MAGIC populations (reviewed by [75]) have been developed for many plant species (Table 1), including many of the world’s most important crops, e.g., rice [76, 77], wheat [78, 79], maize [49], and tomato [80]. The key characteristics of MAGIC populations are the use of multiple founders (typically eight) and multiple rounds of intercrossing, before the development of progeny for genetic mapping. With eight founders, there are 28 possible F_1 (2-way) crosses and 210 possible four-way crosses among unrelated F_1 s. There are then 315 possible ways of creating the eight-way crosses [79] (Fig. 1). Depending on the species, at the four-way stage and eight-way stage, the amount of crossing can become impractical and reduced numbers may be considered. In addition, progeny of four-way crosses are segregating, so there can be an advantage to replicating eight-way crosses with additional four-way parents. The design options and consequences for MAGIC populations have not been fully exhausted, though it is important to maintain a balance of contribution in lines of descent from each founder, to avoid introducing population structure into the population and produce as uniform a decay in LD across the genome as possible.

MAGIC populations afford a number of important benefits over the more commonly used bi-parental and/or association-mapping populations: (1) using multiple parent samples more genetic variation than in any traditional bi-parent cross. (2) The allele frequencies are balanced, because founders contribute equally to the population. (3) Dense, evenly distributed recombination sites provide considerable resolution for genetic analysis, genetic map construction, and gene isolation. MAGIC will work well in species where LD is extensive (such as inbreeding species like rice and wheat), and where LD mapping approaches may not give adequate precision, thus requiring more highly recombined resources. Combined

Table 1 Examples of published plant multi-founder populations

Population type	Species	Founder number	Population details	Genotype data	Reference
NAM	Maize	26	25 ILs crossed to 1 reference line, 5,000 RILs	1.5k SNPs	McMullen et al. [49]
NAM	Barley	26	25 wild barleys crossed to 1 reference line, 1,420 RILs	5.7k SNPs	Maurer et al. [60]
		26	25 wild barleys crossed to 1 reference line, 796 RILs	384 SNPs ^a	Nice et al. [64]
NAM	Bread wheat	11	10 varieties crossed to 1 reference line, 852 RILs	13.4k GbyS-derived SNPs	Bajgain et al. [61]
MAGIC	Bread wheat	4	2 rounds of intercrossing, 1,579 RILs	1,670 DAR/SSR/SNPs	Huang et al. [78]
		8	3 rounds of intercrossing, 1,000 RILs	82k SNP array	Mackay et al. [79]
MAGIC	Tomato	8	3 rounds of intercrossing, 397 RILs	1.5k SNP array	Pascual et al. [80]
MAGIC	Rice	8	ssp. <i>indica</i> . 3 rounds of intercrossing, 1,328 S7 RILs	GbS	Bandillo et al. [76]
		8	ssp. <i>indica</i> . 5 rounds of intercrossing		Bandillo et al. [76]
		8	ssp. <i>japonica</i> . 3 rounds of intercrossing, 500 S5 RILs		Bandillo et al. [76]
		16	4 rounds of intercrossing, ssp. <i>japonica</i> and <i>indica</i>		Bandillo et al. [76]
		4	ssp. <i>indica</i> . 2 rounds intercrossing, 271 RILs	6k SNP array	Meng et al. [77]
		4	ssp. <i>indica</i> . 2 rounds intercrossing, 268 RILs	6k SNP array	Meng et al. [77]
		8	ssp. <i>indica</i> . 3 rounds intercrossing, 531 RILs	6k SNP array	Meng et al. [77]
MAGIC	Barley	8	3 rounds intercrossing, 533 DH lines	4.5k SNPs	Sannemann et al. [90]
MAGIC	<i>Arabidopsis</i>	19	527 RILs	1.3k SNPs	Kover et al. [91]

(continued)

Table 1 (continued)

Population type	Species	Founder number	Population details	Genotype data	Reference
MAGIC	Maize	8	3 rounds of intercrossing, 1,636 RILs	54k SNPs	Dell'Acqua et al. [92]
AMPRIL	<i>Arabidopsis</i>	8	2 rounds of intercrossing, 532 RILs	321 SSR/SNPs	Huang et al. [87]

SNP single nucleotide polymorphism, *DArT* Diversity Array Technology, *SSR* simple sequence repeat, *RIL* recombinant inbred line, *IL* inbred line, *GbyS* genotyping-by-sequencing, *Arabidopsis thaliana*, *Barley Hordeum vulgare*, *Bread wheat Triticum aestivum*, *Maize Zea mays*, *Rice Oryza sativa*, *Tomato Solanum lycopersicum*, Genotype details refer to those listed in the original publication; some populations have been overlaid with additional genotype data subsequently

^aAdditionally, 4,022 SNPs and 263,531 sequence variants imputed into the population via additional genotyping of the parents

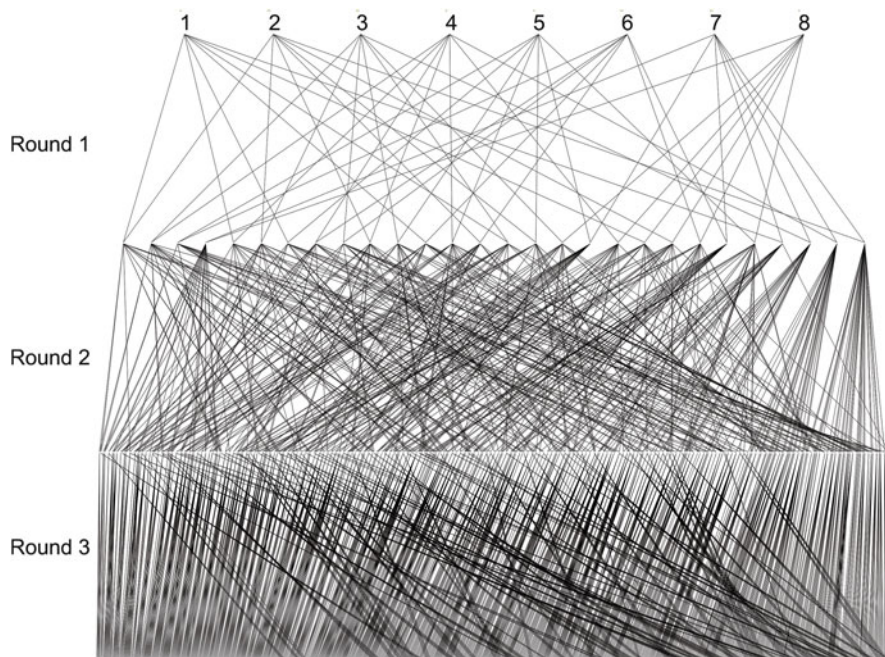


Fig. 1 A balanced eight-way MAGIC crossing scheme. Top line: eight founder parents. Second line: 28 two-way crosses resulting from all pairwise crosses (half-diallel) between the founders. Third line: 210 four-way crosses, resulting from intercrossing all unrelated two-ways. Fourth line: 315 eight-way crosses resulting from intercrossing all unrelated four-ways

with their suitability for the generation of high-density genetic maps, MAGIC populations are ideal community-based resources for crop improvement, fine mapping QTL, QTL \times environment and epistatic effects, and the anchoring of physical-genetic maps. However, a significant disadvantage of MAGIC populations, compared to say a bi-parental population, is the time they take to create. For example, the eight-way winter wheat MAGIC population created by Mackay et al. [79] took 3.5 years to complete: 1.5 years for the intercrossing, plus 2 additional years to reach the F₅ stage of inbreeding. Creation of DH lines from the final outcrossed population would reduce time, but a small number of DH lines are required from a large number of outcrossed individuals, unlike the case in a bi-parental population, where a large number of lines are required from a single individual (the F₁). This makes DH production to generate MAGIC populations prohibitively expensive for species like wheat.

An extension of MAGIC that has yet to be widely tested is the extraction of lines after further generations of outcrossing of the initial population. This was proposed by Mackay and Powell [25]. These lines would be more highly recombined than the initial set and therefore give finer location of QTL. High-density genotyping or sequencing of QTL containing regions identified in the initial population followed by selective phenotyping of recombinants should increase precision for only a modest cost. There is a limit to the number of additional generations of additional crossing that are worth carrying out: the reduction of linkage disequilibrium through crossing is countered by its increase as a result of finite population size together with loss of variation through drift. This process is under development in the 'NIAB Elite MAGIC' (eight-founders) and 'NIAB Diverse MAGIC' (16 founders) populations: an additional four cycles of crossing have been carried out and inbred lines are being created from the more advanced generation.

14 Analysis Approaches for MAGIC

There is a basic split of methods of analysis into those that regress traits onto individual markers and those that regress traits on probabilities of inheritance from founders. Within each of these categories there are then several variants. We give an outline of the three most common below. A fuller account and description are given in Verbyla et al. [81].

Standard statistical methods of analysis to test for marker-trait associations, such as a t-test, do not take into account the method of construction of MAGIC populations, and will result in an increased frequency of false-positive results. A typical MAGIC population is constructed from a set of funnel crosses in which founders are crossed in different orders. For an eight-founder population, there are 315 possible ways in which all founders can be combined. For any polygenic trait, the additive genetic variance expressed between funnels is expected to be 1/8th of the whole. In addition, (1) replicate crosses may be made when selecting parents from four-way and subsequent segregating generations, (2) replicate individuals

within the final generation of crosses may be selected to initiate inbreeding, and (3) more than one inbred line may be derived from each outcrossed individual. These nested relationships among the resulting lines must be taken into account during analysis. For example, we have used gene dropping, as implemented in the bespoke software GeneDrop [82] (available from <http://www.niab.com/>), to simulate a quantitative trait segregating in the “NIAB Elite MAGIC” population. We simulated an eight-founder population with an average of six inbred lines per funnel; 10,000 pairs of unlinked SNPs were tested for association. In the absence of correction for family structure, false-positive rates were 0.1678, 0.0322, 0.0056, and 0.0009 at nominal significance thresholds of 0.1, 0.01, 0.001, and 0.0001, respectively: a considerable increase over expected, particularly at more stringent significance thresholds. Incorporating family structure, the corresponding false-positive rates were 0.1056, 0.0126, 0.0011, and 0.0000: much closer to the nominal significance levels.

The hierarchical structure described above is most easily taken into account in a mixed model incorporating random effects (i.e., variance components) for differences at each level. In a typical eight-founder population, this would include terms for differences between funnels, between replicate crosses within funnels, and between replicate plants within replicate crosses. The marker effects of greatest interest are included as fixed effects. For bi-allelic markers a simple additive model is most powerful, in a one degree of freedom (df) test in which the trait is regressed on the gene dosage (0, 1, 2 in a diploid), of an arbitrarily chosen reference allele. A two df test can also be used in which the three genotype classes are treated separately. However, the heterozygous class is usually rare and, even for dominant QTL, the two df test generally has reduced power. The effect of the heterozygotes can readily be examined after the initial scan, however. Multi-allelic loci are best tested for linkage in a (no. of alleles – 1) test.

In this model, there is no requirement to estimate and incorporate into the analysis a genetic relationship matrix among SSD lines in the manner required in association mapping. An advantage of this approach is therefore that it is easy to implement in almost all standard statistical packages. In R [83] for example, the lmer package [84] can be used to incorporate the desired random effects. As a result, modelling multiple markers, their interactions, and other covariates is also straightforward.

Incorporation of a marker-based relationship matrix is also possible, however, and has been used in mapping within a rice MAGIC population [76]. It has the advantage that relationships among the founder lines are also taken into account, though computational ease and simplicity is reduced.

Single-marker methods of analysis should always be included: they are quick, flexible, and robust to genotype error, which will generally reduce power but not increase the false-positive rate. Alternative methods of analysis use the marker data to estimate probabilities of identity by descent between each RIL and each founder at all selected locations over the genome. Ideally, these probabilities will be one or zero, indicating that a particular location in a line is known to have originated with certainty from one of the founders. In the worst case, these probabilities would be

1/(number of founders). Software to calculate these probabilities is available in *r/qtl* [5], *mpMap* [85], *RABBIT* [86], and *HAPPY* [70]. The first three packages require the pedigree of each line (with reference to the founders) to be known, whereas the latter does not. We are not aware of an independent comprehensive analysis of the absolute and relative accuracies of these methods, while also taking into account their availability, reliability, and ease of use.

Once calculated, traits can be regressed on each founder probability to give a test with 1 df for a QTL allele carried by that founder. Since there would be (no. of founders -1) such independent tests, a multiple regression is carried out on the identity by descent (IBD) values to give a single test with (no. of founders -1) degrees of freedom. To achieve this, one of the founder probabilities (it doesn't matter which) is dropped. Just as for single-marker analyses described above, the hierarchical population structure of the MAGIC populations should still be taken into account. *HAPPY* does not do this, but estimates empirical significance thresholds through a resampling procedure. IBD probabilities can also be calculated for locations between markers. These too, can be used for analysis. This is analogous to interval mapping in bi-parental crosses. IBD methods are generally restricted to additive models: with eight founders there are 28 heterozygous combinations, so the locus specific test for association would require 35 df, assuming all heterozygous classes were represented. The IBD approach to analyzing MAGIC populations allows each founder to carry separate QTL alleles. This will be an advantage over the single (bi-allelic) marker approach in some circumstances. This depends on the true number of QTL alleles, the distribution of their effects, the pattern of LD between marker and QTL alleles, the accuracy of IBD determination, the accuracy of genotyping (which will disproportionately affect IBD probability estimation), and marker density. It is possible that in the near future, IBD methods may be superseded as methods of genotyping and sequencing result in marker densities approaching the limit of capturing all variants segregating in the population (though this will increase the problem of multiple-testing). Our best advice would be to try an IBD-based method and a single-marker method. If results agree, all well and good (and in our experience, they usually do). Lack of agreement should be explored further to establish the cause. More complex models (reviewed by Verbyla et al. [81]) involve approaches analogous to composite interval mapping, Bayesian methods, and can fit multiple QTL models with simultaneous analysis of phenotypes. Sadly, most methods are currently not easily accessible to the non-statistician or data analyst.

14.1 *Arabidopsis* Multi-Parent Recombinant Inbred Line (AMPRIL)

The *Arabidopsis* multi-parent recombinant inbred line (AMPRIL) population described by Huang et al. [87] was developed from eight inbred *Arabidopsis*

accessions from diverse geographical origins. Four unrelated F_1 combinations were made among the eight founders. These F_1 s were crossed in a diallel to give six four-way crosses (pooling reciprocals), which were then selfed to produce 532 inbred lines. This pattern of construction is similar to that for MAGIC, but required fewer crosses and generations to create than the equivalent eight-founder MAGIC population. The comments above about MAGIC populations therefore apply to AMPRIL too. For an equivalent population size, a MAGIC population will provide more resolution than AMPRIL, although AMPRIL would be quicker to create.

14.2 Linked or Multiple Mapping Populations

In principle, any set of mapping populations can be analyzed simultaneously to detect QTL and, by increasing sample size, to increase power and precision. If links between populations can be made, power may be increased further by incorporating information on linkage disequilibrium across populations in addition to linkage information within populations. This has resulted in the use of lines derived from various sets of linked crosses, such as from diallels [88], though these links are not an absolute requirement. The focus of these approaches has been largely on detection of QTL in different genetic backgrounds and on epistasis rather than primarily on improving precision, though that should be a consequence of increasing population size. A recent example, with references to earlier work, is that of Han et al. [89].

15 Conclusion and Outlook

Precision mapping requires that marker-trait associations are tested in populations in which large numbers of recombinations have occurred. To achieve this goal, there are two broad approaches: increase population size and increase the number of generations of crossing. The methods described in this review attempt one or both of these. We favor approaches such as MAGIC and AMPRIL, since these focus explicitly on increasing the amount of recombination that occurs within the population. This bias may be because our own background is of working with inbred crops where LD generally decays quite slowly in collections of lines and the number of elite cultivars available is limited. Consequently, the power and precision of association mapping may be limited. In contrast, approaches that rely on linked sets of crosses, such as NAM, may be better suited to outcrossed species such as maize (where the limitation can be that LD decays too quickly in association-mapping panels, so power is limited but precision is increased). In these circumstances, experimental populations may be required to increase power as much as to increase precision. An equivalent way of viewing the choice would be that MAGIC populations are better for fine mapping in germplasm of immediate relevance to

breeders' elite germplasm, but NAM is better for progression towards gene discovery and positional cloning, since greater diversity can be captured in very diverse germplasm and the use of linkage in addition to linkage disequilibrium protects against loss of power for QTL detection.

Whatever approach is followed, the days of mapping QTL in small populations must come to an end. In our own work in MAGIC wheat populations, we started with a target of developing 1,000 lines per population: that number now looks on the low side.

Acknowledgements JC and IM were partially funded by grants from the Biotechnology and Biological Sciences Research Council (BB/M008908/1, BB/M011666/1 and BB/L011700/1) and the Agriculture and Horticulture Development Board (RD2200003).

Glossary

Advanced inter-cross (AIC) A bi-parental population, in which founders have been intercrossed for two or more generations prior to the production of inbred lines.

Doubled haploid (DH) A genotype formed when haploid cells undergo chromosome doubling.

Genomic selection (GS) A form of marker-assisted selection in which genetic markers are combined with phenotypic data to estimate breeding values in the absence of precise knowledge of where specific genes are located.

Genome wide association scan (GWAS) Method for genetic mapping using a collection of varieties or landraces with phenotypic and genome-wide genotypic datasets.

Linkage disequilibrium (LD) The non-random association of alleles at separate loci located on the same chromosome.

Multiparent advanced generation inter cross (MAGIC) population A multi-founder population created by intercrossing the founders over multiple generations in a balanced crossing scheme, prior to the production of inbred lines.

Nested association mapping (NAM) population A multi-founder population created by generating multiple bi-parental inbred populations, each of which contains a common founder.

Quantitative trait locus (QTL) A polymorphic site contributing to the genetic variability of a quantitative trait.

Recombinant inbred line (RIL) A population developed by single seed descent from the F₂ generation.

References

1. Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8:552
2. Beckmann JS, Soller M (1983) Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. *Theor Appl Genet* 67:35–43
3. Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335:721–726
4. Paux E, Sourdille P, Mackay I, Feuillet C (2012) Sequence-based marker development in wheat: advances and applications to breeding. *Biotechnol Adv* 30:1071–1088
5. Broman KW, Wu H, Sen Ś, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19:889–890
6. Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* 363:557–572
7. Beavis WD (1998) QTL analysis: power, precision, and accuracy. In: Paterson AH (ed) *Molecular dissection of complex traits*. CRC Press, Boca Raton, pp 145–173
8. Kearsey MJ, Farquhar AG (1998) QTL analysis in plants; where are we now? *Heredity* 80:137–142
9. Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci* 48:1649–1664
10. Monna L, Lin HX, Kojima S, Sasaki T, Yano M (2002) Genetic dissection of a genomic region for a quantitative trait locus, *Hd3*, into two loci, *Hd3a* and *Hd3b*, controlling heading date in rice. *Theor Appl Genet* 104:722–778
11. Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for genomic selection applications and research in plants. *Crop Sci* 55(1):12
12. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
13. Mackay I, Ober E, Hickey J (2015) GplusE: beyond genomic selection. *Food Energy Secur* 4:25–35
14. Bortesi L, Fischer R (2015) The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol Adv* 33:41–52
15. Hickey JM, Bruce C, Whitelaw A, Gorjanc G (2016) Promotion of alleles by genome editing in livestock breeding programmes. *J Anim Breed Genet* 133:83–84
16. Law CN, Worland AJ, Giorgi B (1976) The genetic control of ear-emergence time by chromosome 5A and 5D of wheat. *Heredity* 36:49–58
17. Price AH (2006) Believe it or not, QTLs are accurate! *Trends Plant Sci* 11:213–216
18. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14:365–376
19. Nature Genetics Editorial Board (2005) Framework for a fully powered risk engine. *Nat Genet* 37:1153
20. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9:356–369
21. Hall D, Tegström C, Ingvarsson PK (2010) Using association mapping to dissect the genetic basis of complex traits in plants. *Brief Funct Genomics* 9:157–165
22. Bentley AR, Scutari M, Gosman N, Faure S, Bedford F, Howell P, Cockram J, Rose GA, Barber T, Irigoyen J, Horsnell R, Pumfrey C, Winnie E, Schacht J, Beauchêne K, Praud S, Greenland A, Balding D, Mackay IJ (2014) Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. *Theor Appl Genet* 127:2619–2633

23. Cockram J, White J, Zuluaga DL, Smith D, Comadran J et al (2010) Genome-wide association mapping to candidate polymorphism resolution in the un-sequenced barley genome. *Proc Natl Acad Sci U S A* 107:21611–21616
24. Waugh R, Marshall D, Thomas B, Comadran J, Russell J, Close T, Stein N, Hayes P, Muehlbauer G, Cockram J, O’Sullivan D, Mackay I, Flavell A, Agoueb A, Barley CAP, Ramsay L (2010) Whole-genome association mapping in elite inbred crop varieties. *Genome* 53:967–972
25. Mackay IJ, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
26. MacArthur D (2012) Methods: face up to false positives. *Nature* 487:427–428
27. Highfill CA, Reeves GA, Macdonald SJ (2016) Genetic analysis of variation in lifespan using a multiparental advanced intercross Drosophila mapping population. *BMC Genet* 17(1):113
28. Cockram J, White J, Leigh FJ, Lea VJ et al (2008) Association mapping of partitioning loci in barley (*Hordeum vulgare* ssp. *vulgare* L.) *BMC Genet* 9:16
29. Darvasi A, Soller M (1995) Advanced intercross lines, and experimental population for fine genetic mapping. *Genetics* 141:1199–1207
30. Ma J, Wingen LU, Orford S, Fenwick P, Wang J, Griffiths S (2015) Using the UK reference population Avalon x Cadenza as a platform to compare breeding strategies in elite Western European bread wheat. *Mol Breed* 35:70
31. Bentley AR, Jensen EF, Mackay IJ, Hönicka H, Fladung M, Hori K, Yano M, Mullett JE, Armstead IP, Hayes C, Thorogood D, Lovatt A, Morris R, Pullen N, Mutasa-Göttgens E, Cockram J (2013) Genomics and breeding for climate-resilient crops (ed Kole C) volume II target traits chapter 1. Flowering time. Springer, Berlin
32. Bentley A, Mackay I (2016) Advances in wheat breeding techniques. In: Langridge P (ed) Achieving sustainable cultivation of wheat. Burleigh Dodds Science Publishing Ltd., Cambridge
33. Zou C, Wang P, Xu Y (2016) Bulk sample analysis in genetics, genomics and crop improvement. *Plant Biotechnol J* 14:1941–1955
34. Routaboul J-M, Dubos C, Beck G, Marquis C, Bidzinski P, Loudet O, Lepiniec L (2012) Metabolite profiling and quantitative genetics of natural variation for flavonoids in Arabidopsis. *J Exp Bot* 63:3749–3764
35. Ries D, Holtgräwe D, Viehöver P, Weisshaar B (2016) Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. *BMC Genomics* 17:236
36. Hill WG (1998) A note on the theory of artificial selection in finite populations and application to QTL detection by bulk segregant analysis. *Genet Res* 72:55–58
37. Mackay IJ, Caligari PDS (2000) Efficiencies in F₂ and backcross generations for bulked segregant analysis using dominant markers. *Crop Sci* 40:626–630
38. Fitz Gerald JN, Carlson AL, Smith E, Maloof JN, Weigel D, Chory J, Borevitz JO, Swanson RJ (2014) New Arabidopsis advanced intercross recombinant inbred lines reveal female control of nonrandom mating. *Plant Physiol* 165:175–185
39. Balint-Kurti PJ, Wisser R, Zwonitzer JC (2008) Use of an advanced intercross line population for precise mapping of quantitative trait loci for gray leaf spot resistance in maize. *Crop Sci* 48:1696–1704
40. Balint-Kurti PJ, Zwonitzer J, Wisser R (2008) Use of an advanced intercross line population for precise mapping of quantitative trait loci for grey leaf spot resistance in maize. *Crop Sci* 48:1696–1703
41. Kooke R, Wijker E, Keurentjes JJ (2012) Backcross populations and near isogenic lines. *Methods Mol Biol* 871:3–16
42. Fletcher RS, Mullen JL, Yoder S, Bauerle WL, Reuning G, Sen S, Meyer E, Juenger TE, McKay JK (2013) Development of a next-generation NIL library in *Arabidopsis thaliana* for dissecting complex traits. *BMC Genomics* 14:655
43. Gale JS (1980) Population genetics. Blackie and Son, Glasgow and London

44. Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. *Theor Appl Genet* 95:1005–1011
45. Yamanaka N, Watanabe S, Toda K, Hayashi M, Fuchigami H, Takahashi R, Harada K (2005) Fine mapping of the *FT1* locus for soybean flowering time using a residual heterozygous line derived from a recombinant inbred line. *Theor Appl Genet* 110:634–639
46. Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539–551
47. Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279–292
48. Guo B, Sleper DA, Beavis WD (2010) Nested association mapping for identification of functional markers. *Genetics* 186:373–383
49. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Lu H et al (2009) Genetic properties of a maize nested association mapping population. *Science* 178:539–551
50. Brown PJ, Upadaya N, Mahone GS, Tian F, Bradbury PJ et al (2011) Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet* 7:e1002383
51. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ et al (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
52. Hung H-Y, Shannon LM, Tian F, Bradbury PJ, Chen C et al (2012) *ZmCCT* and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. *Proc Natl Acad Sci U S A* 109:E1913–E1921
53. Kump KL, Bradbury PJ, Wissner RJ, Buckler ES, Belcher AR et al (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet* 43:163–168
54. Peiffer JA, Flint-Garcia SA, De Leon N, McMullen MD, Kaeppler SM et al (2013) The genetic architecture of maize stalk strength. *PLoS One* 8:e67066
55. Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z et al (2014) The genetic architecture of maize height. *Genetics* 196:1337–1356
56. Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc Natl Acad Sci U S A* 108:6893–6898
57. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162
58. Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES (2014) Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet* 10:e1004845
59. Jordan D, Mace E, Cruickshank A, Hunt C, Henzell R (2011) Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. *Crop Sci* 51:1444–1457
60. Maurer A, Draba V, Jiang Y, Schnaithmann F, Sharma R, Schumann E, Killian B, Reif JC, Pillen K (2015) Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16:290
61. Bajgain P, Rouse MN, Tsilo TJ, Macharia GK, Bhavani S, Jin Y, Anderson JA (2016) Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. *PLoS One* 11:e0155760
62. Wingen LU, West C, Leverington-Waite M, Collier S, Orford S et al (2017) Wheat landrace genome diversity. *Genetics* 205:1657–1676
63. Stich B (2009) Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. *Genetics* 183:1525–1534
64. Nice LM, Steffenson BJ, Brown-Guedira GL, Akhunov ED, Liu C, Kono TJY, Morrell PL, Blake TK, Horsley RD, Smith KP, Meuhlbauer GJ (2016) Development and genetic characterization of an advanced backcross-nested association mapping (AB-NAM) population of wild x cultivated barley. *Genetics* 203:1453–1467
65. Moore G (2015) Strategic pre-breeding for wheat improvement. *Nat Plants* 1:15018

66. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2202
67. Tversky A, Kahneman D (1971) Belief in the law of small numbers. *Psychol Bull* 76:105
68. 3000 Rice Genomes Project (2014) The 3000 rice genomes project. *Gigascience* 3:7
69. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* 43:956–963
70. Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci U S A* 97:12649–12654
71. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JNP, Mott R, Flint J (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* 38:879–887
72. Goldringer I, Enjalbert J, David J, Paillard S, Pham JL et al (2001) Dynamic management of genetic resources: a 13-year experiment on wheat. In: Cooper HD, Spillane C, Hodgkin T (eds) *Broadening the genetic base of crop production*. CABI, Wallingford, pp 245–260
73. Thépot S, Restoux G, Goldringer I, Gouache D, Mackay I, Enjalbert J (2015) Efficiently tracking selection in a multiparental population: the case of earliness in wheat. *Genetics* 199:609–623
74. The Complex Trait Consortium (2002) The collaborative cross, a community resource for the genetic analysis of complex traits. *Nat Genet* 36:1133–1137
75. Huang BE, Verbyla KL, Verbyla AP, Raghavan C, Singh VK, Gaur P, Leung H, Varshney RK, Cavanagh CR (2015) MAGIC populations in crops: current status and future prospects. *Theor Appl Genet* 128:999–1017
76. Bandillo N, Raghavan C, Muyca PA, Sevilla MAL, Lobina IT (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetic research and breeding. *Rice* 6:11
77. Meng L, Guo L, Ponce K, Zhao X, Ye G (2016) Characterization of three *indica* rice multiparent advanced generation intercross (MAGIC) populations for quantitative trait loci identification. *Plant Genome* 9(2).
78. Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012) A multiparent advanced generation inter-cross population for genetic analysis of wheat. *Plant Biotechnol J* 10:826–839
79. Mackay I, Bansept-Basler P, Barber T, Bentley AR, Cockram J et al (2014) An eight-parent multiparent advanced generation intercross population for winter-sown wheat: creation, properties and validation. *G3 (Bethesda)* 4:1603–1610
80. Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet JP, Le QH, Chauchard B, Verschave P, Causse M (2015) Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnol J* 13:565–577
81. Verbyla AP, George AW, Cavanagh CR, Verbyla KL (2014) Whole-genome QTL analysis for MAGIC. *Theor Appl Genet* 127:1753–1770
82. Ladejobi O, Elderfield J, Gardner KA, Gaynor RC, Hickey J, Hibberd JM, Mackay IJ, Bentley AR (2016) Maximizing the potential of multi-parental crop populations. *App Transl Genom* 11:9–17
83. R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL: <https://www.R-project.org/>
84. Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *J Stat Softw* 67:1–48
85. Huang BE, George AW (2011) R/mpMap: a computational platform for the genetic analysis of multiparent recombinant inbred lines. *Bioinformatics* 27:727–729

86. Zheng C, Boer MP, van Eeuwijk F (2015) Reconstruction of genome ancestry blocks in multiparental populations. *Genetics* 200:1073–1087
87. Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA (2011) Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. *Proc Natl Acad Sci U S A* 108:4488–4493
88. Rebai A, Goffinet B (1993) Power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theor Appl Genet* 86:1014–1022
89. Han S, Utz HF, Liu W, Schrag TA, Stange M, Würschum T, Miedaner T, Bauer E, Schön CC, Melchinger AE (2016) Choice of models for QTL mapping with multiple families and design of the training set for prediction of Fusarium resistance traits in maize. *Theor Appl Genet* 129:431–444
90. Sannemann W, Huang BE, Mathew B, Léon J (2015) Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. *Mol Breed* 35:86
91. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet* 5:e7
92. Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G, Hlaing AL, Aung HH, Nelissen H, Baute J, Frascaroli E, Churchill GA, Inzé D, Morgante M, Pé ME (2015) Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biol* 16:167

TILLING: The Next Generation

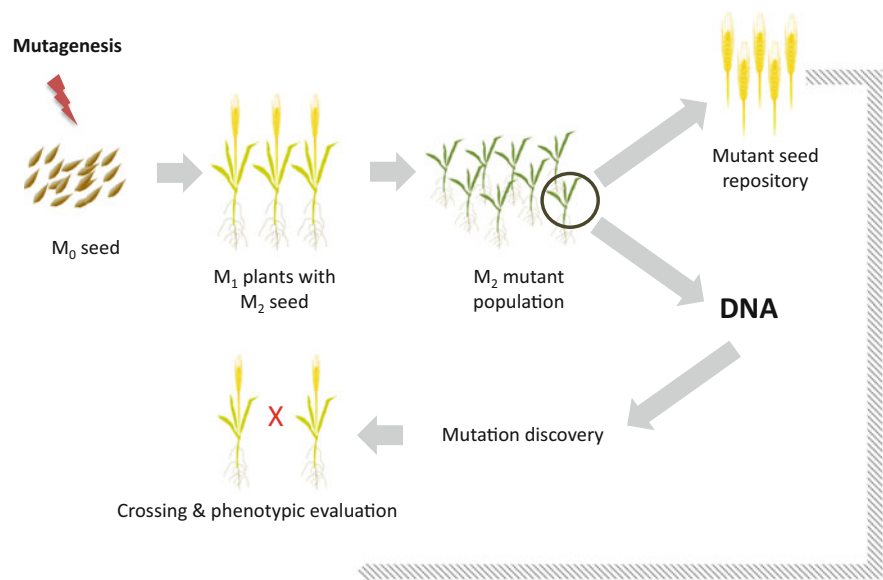


Bradley J. Till, Sneha Datta, and Joanna Jankowicz-Cieslak

Abstract Gene space: the final frontier in plant functional genomics. These are the voyages of TILLING, the reverse-genetics strategy that sought to boldly go where no-one had gone before by combining high-density chemical mutagenesis with high-throughput mutation discovery. Its 18-year mission has been to explore new technologies such as next generation sequencing and to seek out new strategies like in silico databases of catalogued EMS-induced mutations from entire mutant plant populations. This chapter is a clip show highlighting key milestones in the development of TILLING. Use of different technologies for the discovery of induced mutations, establishment of TILLING in different plant species, what has been learned about the effect of chemical mutagens on the plant genome, development of exome capture sequencing in wheat, and a look to the future of reverse-genetics with targeted genome editing are discussed.

B. J. Till (✉), S. Datta, and J. Jankowicz-Cieslak
Plant Breeding and Genetics Laboratory, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, IAEA Laboratories Seibersdorf, International Atomic Energy Agency, Vienna International Centre, Vienna, Austria
e-mail: b.till@iaea.org

Graphical Abstract



Keywords Chemical mutagenesis, CRISPR/Cas, EMS, In silico TILLING, Next generation sequencing

Contents

1 Introduction 141

2 The First TILLING Service and Expansion into Other Plant Species 146

3 Next-Generation TILLING 148

4 Towards In Silico TILLING 151

5 Reverse-Genetics Using Targeted Genome Editing 153

6 Choosing the Best Approach 154

7 Concluding Remarks and Future Perspectives 155

References 156

Symbol

φ Greek letter Phi

Abbreviations

2X Diploid
 4X Tetraploid

6X	Hexaploid
ATP	Arabidopsis TILLING Project
Az	Azide
CIAT	International Center for Tropical Agriculture
CRISPR	Clustered Regularly-Interspaced Short Palindromic Repeats
CRISPRa	CRISPR activator
CRISPRi	CRISPR interference
DNA	Deoxyribonucleic acid
DSBs	Double strand breaks
EMC	Enzymatic mismatch cleavage
EMCA	EMC with agarose gel
EMCC	EMC with capillary electrophoresis
EMCH	EMC with HPLC
EMCL	EMC with LI-COR gels
EMCP	EMC with polyacrylamide gels
EMS	Ethyl methanesulfonate
ENU	<i>N</i> -ethyl- <i>N</i> -nitrosourea
Gb	Giga bases
HDR	Homology-directed repair
HPLC	High performance liquid chromatography
HRM	High resolution melt
indel	Insertion or deletion of bases
kb	Kilobases
M ₀	Plant generation prior to mutagenesis
M ₁	First generation of mutagenized plant
M ₂	Second generation of mutagenized plant
Mbp	Million base pairs
MNU	<i>N</i> -Nitroso- <i>N</i> -methylurea
NHEJ	Non-homologous end joining
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
sgRNA	Single guide RNA
SNP	Single nucleotide polymorphism
TALENs	Transcription Activator-Like Effector-based Nucleases
TILLING	Targeting Induced Local Lesions IN Genomes
ZFNs	Zinc finger nucleases

1 Introduction

The Dutch botanist Hugo de Vries is credited as the first person to introduce the word mutation to the scientific vocabulary. His “mutation theory” was based in part on observations of spontaneous and heritable phenotypic changes (mutations) occurring

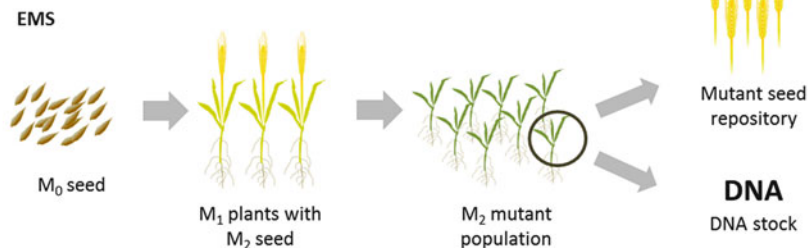
in evening primroses over a 13-year period [1]. What de Vries was observing was later determined to be the result of large chromosomal aberrations unique to *Oenothera* species. It was the work of Thomas Hunt Morgan and colleagues in the first quarter of the twentieth century on *Drosophila melanogaster* that would popularize the use of the word “mutation” to describe genetic variations in single genes [2]. In addition to stimulating mutation research, de Vries would later go on to describe the phenomenon of genetic recombination in 1903 [3]. Thus, by the early 1900s the major driving forces of genetic diversity, mutation and recombination, were described. These two events underlie biological evolution and provide the means for humans to generate novel diversity in plants and animals (Fig. 1).

Mutations are a particularly useful tool for both geneticist and breeder. New mutations create novel alleles that can have a profound impact on organismal phenotype, and provide the raw material for breeders to create combinations of alleles to improve crop performance [5]. While spontaneous mutations are a major source of heritable phenotypic diversity, they pose a problem for the researcher: they happen quite rarely. Indeed, recent studies employing whole genome sequencing suggest a spontaneous mutation rate of 7.4×10^{-9} in rice and 7×10^{-9} in *Arabidopsis* [6, 7]. A major milestone, therefore, was the discovery that mutations could be induced much faster than they appear in nature.

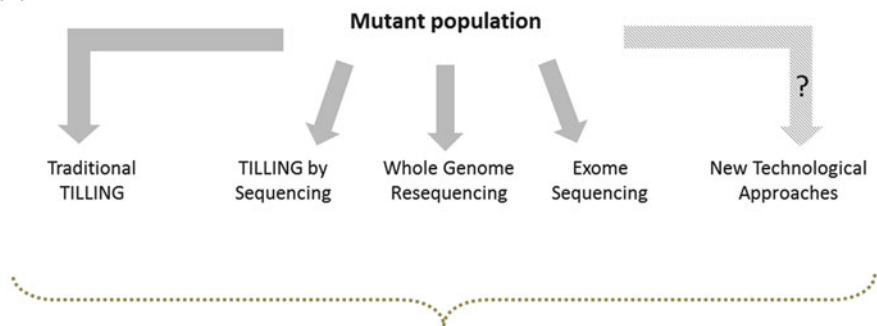
Herman Muller used X-rays to create mutations in *Drosophila melanogaster* that accumulated orders of magnitude faster than what was observed spontaneously [8]. Contemporary with this, Lewis John Stadler used X-rays to induce mutations in cereals [9, 10]. The idea that mutations could be used for breeding was quickly adopted and by the late 1930s the first mutant crop variety was released, a cultivar of tobacco named Chlorina that had improved characteristics for cigar smoking [11, 12]. This ushered in the field of plant mutation breeding that has resulted in the official release of more than 3,200 mutant crop varieties [13, 14]. Forward genetic approaches that utilize induced mutations remain popular likely because of the ease of mutation induction in many crops and the fact that phenotypes can be observed without any prior knowledge of genes or gene function.

Activities to determine the sequence of DNA, and thus genes, in organisms began in the 1960s and led to the first full DNA genome (bacteriophage ϕ X) in 1977 [15]. Years later, the development of next-generation sequencing technologies has led to a massive increase in the acquisition of gene sequences that had vastly outpaced the establishment of *in vivo* functions of genes through direct experimental evidence. Reverse-genetic methods can bridge this gap as they provide direct *in vivo* testing of the function of genes. The process involves the creation of gene disruptions in the selected genotype, the identification of individuals having affected gene sequences or gene expression, and the testing of these organisms to determine the phenotypic consequence of the mutation (Fig. 2). This is in essence the opposite direction of traditional genetic analysis, where plants are selected based on phenotype and only later are analyzed to determine the genetic alteration that is causative for the observed trait. Thus, the process is the “reverse” of traditional genetic analysis. A key component of reverse-genetic approaches is that they are hypothesis driven endeavors where the researcher seeks to study the *in vivo* function of a gene

(A) Population development



(B) Mutation detection



(C) Phenotypic evaluation

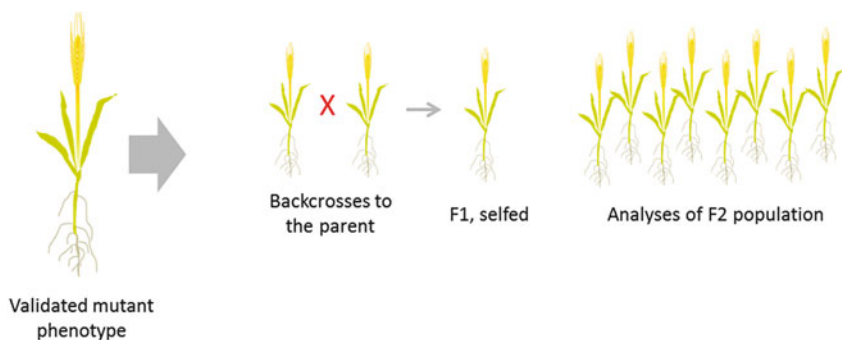


Fig. 2 Overview of the TILLING procedure. The first step is the development of a mutagenized population (a). The chemical mutagen ethyl methanesulfonate (EMS) is typically used. The goal is to obtain a high density of induced point mutations while maintaining suitable survivability and fecundity. While seed mutagenesis is common, examples exist of pollen and tissue culture mutagenesis [16, 17]. For seed propagated crops a single-seed descent strategy is often employed so that the maximum mutation diversity can be captured with the minimum of samples to screen. The optimal population size depends on the density and spectra of induced mutations. Higher mutation densities are achieved in polyploids and thus smaller population sizes are required [18–20]. DNA and seed are collected from plants selected for the TILLING population. The time for the

or other sequence element. Some prior knowledge of appropriate targets (genomic sequences) is therefore required. Candidate targets can often be chosen based on homology to sequences in other organisms where some evidence exists of their function. Prior to the advent of TILLING, reverse-genetics approaches had several limitations including the fact that many were species-specific, using, for example, endogenous transposons, or employed transient disruptions that were not heritable [14].

Chemical mutagenesis was first described in the 1940s with the observation of “chemical production of mutations” in *Drosophila* treated with mustard gas [22]. Mutagens such as EMS became popular and ubiquitous in forward-genetic studies that aimed to elucidate gene function and biological pathways in model organisms. Indeed, many groundbreaking discoveries such as cell cycle control in yeast, segment polarity in *Drosophila*, and meristematic cell signaling in *Arabidopsis* were achieved by forward-genetic screens using EMS [23–25]. Chemical mutagens were thus firmly established by the 1980s as compounds that could produce a high frequency of useful, heritable, and stable mutations for gene function studies. Pioneering work using observed phenotypes provided early estimations on the frequency of genic mutations and optimal population sizes when using EMS [26]. Early work also established that EMS induces primarily point mutations in plants [27].

By the 1990s, technologies for rapid and accurate discovery of SNP variations advanced enough to enable the formulation of reverse-genetics approaches utilizing mutagens inducing primarily single base substitutions. The first reports used denaturing high-performance liquid chromatography (HPLC) for the discovery of point mutations in *Arabidopsis thaliana* and *Drosophila melanogaster* [28, 29]. The *Arabidopsis* group coined the term TILLING (Targeting Induced Local Lesions IN Genomes) for this approach (Fig. 2). This name became widely adopted for subsequent reverse-genetic projects in plants and animals that employed mutagens causing primarily small (SNP and indel) variations [30].

←

Fig. 2 (continued) development of a TILLING population varies and can take more than 1 year for field propagated crops. The second step of TILLING is screening the DNA library for induced mutations (b). Since the inception of TILLING this has been the fastest step. With classical mismatch cleavage and fluorescence detection, an allelic series of ~30-point mutations could be discovered in 1 week using a single DNA analyzer machine [21]. Next-generation sequencing methods have allowed much higher throughputs and the possibility of indexing all mutations in a TILLING population in a short time rather than taking a gene by gene approach. Technologies for DNA sequence evaluation are constantly improving and new approaches will eventually make the discovery and assignment of millions of EMS mutations to individual samples a routine affair. The final step in the TILLING process is testing the effect of the discovered mutations on the mutant plant (c). This step is, and will likely remain, the bottleneck for TILLING or any other reverse-genetic approach. Owing to the high density of background mutations induced by chemical mutagens, one or more backcrosses may be needed to unambiguously correlate genotype with phenotype

2 The First TILLING Service and Expansion into Other Plant Species

Immediately upon the first description of TILLING, efforts were made to improve technologies for mutation discovery so that sample throughput could be increased while at the same time reducing false-positive and false-negative error rates. Development and adaption of mutation discovery technologies for TILLING remains an active area of research as described later in this chapter. A major milestone in the early days of TILLING was the adaptation of enzymatic mismatch cleavage (EMC) for SNP discovery. The activity of single-strand-specific nucleases to cleave single-base-pair mismatches had been reported as early as the 1970s [31, 32], but progress and interpretation of the activity of nucleases on single-base mismatches was hindered due to limitations in available methods to observe cleaved DNA fragments [33, 34]. Henikoff and colleagues developed a method that paired enzymatic mismatch cleavage, eightfold sample pooling, base-pair resolution denaturing polyacrylamide gel electrophoresis, and laser-based fluorescence detection. This approach was termed “high-throughput TILLING” owing to the fact that 768 mutant plants could be screened for mutations in approximately 1 million base pairs in a single gel run [35]. The method proved to be highly robust and accurate and became widely used for mutation discovery in the first decade of TILLING [30, 36].

The major inputs into a TILLING project are the development of suitably a mutagenized population and the generation of a library of high quality genomic DNAs. It was clear from the initiation of the first TILLING efforts in *Arabidopsis thaliana*, that the TILLING population could become a valuable community resource. The first TILLING service was started in 2001 for *Arabidopsis* [37]. Users of the service interfaced online with the *Arabidopsis* TILLING Project (ATP) website. A suite of computational tools guided requestors to choose optimal genic regions of ~1.5 kb to screen for mutations, design PCR primers, and place orders [21]. The ATP would then screen a population of 3,000–6,000 mutagenized lines for mutations in the chosen amplicons, deliver results of alleles discovered, and provide access to seed. In cases where a user requested mutations in a gene that had been previously screened, the requestor was provided a list of mutations already discovered. Thus, within the first year of TILLING being established, one can observe the beginnings of *in silico* TILLING. The ATP later changed its name to the Seattle TILLING Project as it developed a service for TILLING in *Drosophila melanogaster* and collaborated with other groups to expand TILLING into other species such as rice, maize, and soybean [17, 38, 39]. To date, classical TILLING has been reported for over 25 plant species (Table 1 and [36]). TILLING services expanded as other groups provided screening for a range of different plants including rice, tomato, *Brassica rapa*, *Lotus japonicus*, tetraploid and hexaploid wheat, pea, and zebrafish [55–59]. Facilities have either provided screening for free or have charged a fee to recover costs. One issue with single customer-based cost-recovery services is that they depend on having a minimal number of requests over a set period of time to ensure a stable flow of resources to support staff. Sustainability of

Table 1 Selected examples of TILLING projects

Species (common name ^a), ploidy	Mutagen	Mutation frequency 1/kb	Mutation detection technology ^b	References
<i>Arabidopsis thaliana</i> (Arabidopsis ^a), 2X	EMS	1/200	EMCL	[21, 40]
<i>Arabidopsis thaliana</i> (Arabidopsis ^a), 4X	EMS	1/51.5	Illumina amplicon	[41]
<i>Arachis hypogaea</i> L. (peanut), 4X	EMS	1/967	EMCL	[42]
<i>Arachis hypogaea</i> L. (peanut), 4X	EMS	1/344 kb (single copy) 1/3,028 (multi-copy)	Illumina amplicon	[43]
<i>Brassica napus</i> (canola), 2X	EMS	1/109	Illumina amplicon	[44]
<i>Eragrostis tef</i> (tef), 4X	EMS	1/115; 1/370	454 amplicon	[45]
<i>Helianthus annuus</i> L. (sunflower), 2X	EMS	1/475	EMCL	[46]
<i>Helianthus annuus</i> L. (sunflower), 2X	EMS	1/480	EMCL	[47]
<i>Hordeum vulgare</i> (barley), 2X	EMS	1/1,000	EMCH	[48]
<i>Hordeum vulgare</i> (barley), 2X	EMS	1/500	EMCL	[49]
<i>Hordeum vulgare</i> (barley), 2X	EMS	1/1,333	454 amplicon	[50]
<i>Musa acuminata</i> (banana), 3X	EMS	1/57	EMCL	[16]
<i>Oryza sativa ssp. japonica</i> (rice ^a), 2X	EMS Az- MNU	1/294 1/265	EMCL Illumina amplicon, exome capture/ Illumina	[39, 51, 61]
<i>Oryza sativa ssp. japonica</i> (rice), 2X	MNU	1/135	EMCC	[52]
<i>Triticum aestivum</i> (hexaploid wheat), 6X	EMS	1/24	EMCL	[19]
<i>Triticum aestivum</i> (hexaploid wheat), 6X	EMS	1/38	EMCP Exome capture/ Illumina	[20, 62]
<i>Triticum aestivum</i> (hexaploid wheat), 6X	EMS	1/23.3 to 1/37.5	EMCA	[18]
<i>Triticum aestivum</i> (hexaploid wheat), 6X	EMS	1/34; 1/47	EMCA, EMCP	[53]
<i>Triticum durum</i> (tetraploid wheat), 4X	EMS	1/40	EMCL	[19]
<i>Triticum durum</i> (tetraploid wheat ^a), 4X	EMS	1/51	EMCP, exome capture/Illumina	[20, 62]
<i>Triticum monococcum</i> (diploid wheat), 2X	EMS	1/92	EMCA	[54]
<i>Zea mays</i> (corn ^a), 2X	EMS	1/500	EMCL	[17]

^aIndicates present or former TILLING service

^bEMC (+ symbol) Enzymatic mismatch cleavage using one type of readout platform, A agarose gel, C capillary electrophoresis, H HPLC, L LI-COR, P Polyacrylamide gel

public sector TILLING has thus been an issue and several services have already closed down. Development of fully sequenced TILLING libraries as complete *in silico* resources may be a more sustainable model as it requires only limited labor and resources to maintain databases and seed stocks. This has been possible in recent years through advances in genome sequencing technologies (see below).

One result of the expansion of TILLING into different plant species was the rapid acquisition of data on the effect of chemical mutagens on the plant genome. Keeping in mind that the pre-NGS mutation discovery methods used are highly biased for the recovery of SNP and small indel mutations, data from thousands of discovered EMS mutations showed that for many species the majority of induced changes were G:C to A:T transitions (Table 1, [40]). This supports earlier studies showing EMS alkylating the G residue at the O'6 position resulting in the replication machinery incorporating a T rather than a C in the newly synthesized strand. In some species nearly 100% transition changes have been observed. This deviates in other species, owing possibly to alkylation of other oxygens, variations in DNA repair, and pathways involving depurination [60]. Few mutation hot-spots or regional biases have been reported in studies with data sets large enough to provide statistical significance. Rather, data suggests that EMS results in a generally random distribution of mutations across euchromatic chromosomal locations with some local bias based on adjacent base-pairs [40, 61]. The adoption of next generation sequencing for TILLING screens in the last 5 years has resulted in an increase in datasets on the effect of EMS in plants by two orders of magnitude. The analysis of millions of mutations discovered in wheat will help address the issue of any positional bias in the accumulation of EMS induced changes.

Other chemicals and combinations of chemicals such as sodium azide–MNU have been successfully used for TILLING in plants. Mutation densities reported are similar to that with EMS, while the spectra differ slightly (Table 1). The choice of mutagen may be important in species/genotypes where achieving a high density of mutations with EMS is somehow prohibited due to a cytotoxic barrier or some other effect. Chemical mutagens such as EMS can also result in double strand breaks (DSBs) that could cause larger chromosomal aberrations that were not detected in mutation discovery methods employing PCR amplicons. This is a potentially interesting phenomenon that may be observed when using whole genome or reduced representation genome sequencing approaches. Indeed, analysis suggests that large deletions are induced in polyploid wheat [62]. The frequency of such events is predicted to be quite low compared to SNPs, owing to the fact that large changes will likely be more deleterious, resulting in higher sterility and lower heritability.

3 Next-Generation TILLING

One continual field of study in TILLING has been the development and adaptation of different methods for mutation discovery (Fig. 2). During the first decade of TILLING, numerous publications reported alternative methods for SNP discovery

with the ultimate goal of increasing sensitivities and thus improving throughput and reducing costs. These included capillary and gel-based systems, High Resolution Melt (HRM) analysis, denaturing HPLC coupled with enzymatic mismatch cleavage, conformation-sensitive capillary electrophoresis, and mass spectroscopy [36]. While each method has its advantages and disadvantages, none proved to be such a substantial improvement that it replaced the predominant mode of mutation discovery of enzymatic mismatch cleavage and fluorescence detection. Rather, laboratories adopted the best fit for their purpose based on run-costs, amplicon length, equipment maintenance and automation. This began to change with the commercialization of next generation sequencing. Massively parallel whole genome sequencing coupled with bioinformatics analyses allows rapid discrimination of rare sequence variants versus errors due to the sequencing process [6]. The approach offers a vast improvement on sample screening throughput while dramatically reducing wet bench experiments. Disadvantages include the production of very large data sets, a high bioinformatics load, and higher costs. In addition, much of the cost is spent on sequencing nucleotides outside of genes that will have no phenotypic consequence when mutated.

A natural solution for the discovery of chemically induced mutations using NGS was the adaptation of the original TILLING method of screening PCR amplicons rather than sequencing whole genomes. TILLING remains the same except for the mutation discovery step. Several versions of this have been described (Table 1). All approaches share the goal of maximizing screening throughput by increasing the number of samples screened, the level of pooling, and/or the number of amplicons (total bases of unique sequence) screened. In addition to increasing throughput, sample pooling strategies can also increase the accuracy of mutation calls and allow the determination of the exact individual harboring the identified induced mutation in a pool of samples. Two-dimensional eightfold pooling was used in traditional TILLING screens whereby discovery of a mutation in a row and column pool provided the coordinates of the position of the mutant sample arrayed on a 96-well plate [16]. Higher level pooling is possible with next generation sequencing and so three-dimensional strategies could be considered where samples are arrayed in a cube of stacked plates and mutations are identified in row, column, and plate pools providing the x , y , and z coordinates to identify the exact sample having the mutation. This was used in the TILLING by Sequencing approach described by Comai and colleagues where they screened a total of 768 individual rice mutants in a three-dimensional pool consisting of two dimensions of samples pooled 48-fold and one dimension pooled 64-fold [51]. The group also used TILLING by Sequencing to discover EMS induced mutants in wheat. PCR amplification in this approach closely followed that previously reported for traditional TILLING with single-amplicon reactions performed with pooled genomic DNA [39]. One important issue that was addressed in this work was the fact that higher pooling requires higher amounts of genomic DNA in PCR reactions to ensure that when performing a PCR on a pool, amplification occurs on template DNA from all samples. Failure to achieve this would result in elevated false negative error rates. After PCR products were quantified and then pooled, amplicons were fragmented to an appropriate size for library

preparation, and sequencing was performed using the Illumina platform. Purpose-built bioinformatics tools were also developed for mutation calling and are freely available [63]. Comai and colleagues would also use this method for the development of a tetraploid *Arabidopsis* TILLING population showing a density of 19.4 mutations per Mb [41]. While many different next generation sequencing technologies have been described, Illumina is currently the most popular for TILLING by Sequencing and exome capture TILLING projects. Haughn and colleagues described a modification of the TILLING approach to identify EMS induced mutations in three-dimensionally pooled DNA samples of polyploid canola [44]. A three-dimensional pooling approach using multiplex semi-nested PCR was described for recovery of sodium azide-induced mutations in rice [64]. Ozias-Akins and colleagues used the TILLING by Sequencing approach, employing two-dimensional pooling, to recover mutations in single and multi-copy stress resistance genes in peanut [43]. PCR products were typically fragmented prior to sequencing because amplicon lengths were greater than available sequencing read lengths.

As read lengths have increased with the Illumina platform it is now possible to consider direct sequencing of amplicons without fragmentation. This may be especially efficient in organisms with small exons such as zebrafish. Moens and colleagues described a strategy for direct sequencing of 250 base-pair amplicons using Paired-End sequencing to find induced mutations in *N*-ethyl-*N*-nitrosourea (ENU) mutagenized zebrafish [65]. Similar work is being carried out using 600 base-pair amplicons and 2×300 Paired-End reads to identify EMS induced mutations in tomato [66]. One interesting aspect of the zebrafish work surrounds the type of alleles induced by chemical mutagens. From the start of TILLING, efforts were made to integrate predictions of the effect of point mutations on gene function for optimal primer design and to prioritize identified mutants for phenotypic characterization [67–69]. Owing to the fact that splice-site and nonsense changes are easy to predict, activities surrounded the evaluation of missense changes (where the mutation causes a change from one amino acid to another). In general, only about 5% of EMS induced mutations in an average plant gene will be splice-site or nonsense mutations, and only a fraction of missense changes will be predicted to alter gene function. Therefore, on average more than half of mutations identified in a TILLING screen are expected to be of no value. Why then should efforts be undertaken to identify the individual sample that harbors an unwanted induced mutation? An alternative strategy is to screen larger one-dimensional pools of samples in order to capture all mutations as efficiently as possible. The next step is to evaluate the effect of mutations and choose only those of interest to follow up. This approach was used for zebrafish TILLING. One-dimensional pools of DNA from 288 fish were first screened using the Illumina MiSeq. High throughput genotyping assays (HRM) were then designed for specific genes and all individuals from a pool were screened to identify the one harboring the sought after mutation. A similar approach is being used to identify natural mutations in cassava accessions held at the International Center for Tropical Agriculture (CIAT) and also for TILLING by Sequencing in soybean [70, 71].

4 Towards In Silico TILLING

With advances in next generation sequencing, one can consider developing an in silico resource where all mutations from a mutagenized population are discovered simultaneously and recorded in a database. This is in contrast to traditional TILLING where the user orders mutations in a specific gene prior to screening (Fig. 3).

The in silico TILLING approach allows researchers to get results on available mutations in his or her target gene immediately. The challenge with creating such a resource is that while sequencing costs have reduced, many plant genomes are large and accurate discovery of rare SNP mutations requires a suitable depth of coverage. While examples do exist of whole genome sequencing of thousands of plant accessions in order to uncover natural nucleotide variation, the approach remains cost-prohibitive for most TILLING projects [72, 73]. An alternative way is to sequence only a subset of genomic DNA that is most likely to cause phenotypic variation when mutated (Fig. 4). The first example of this is in zebrafish where DNA was enriched with the annotated exons of all 26,206 protein coding genes [74]. This

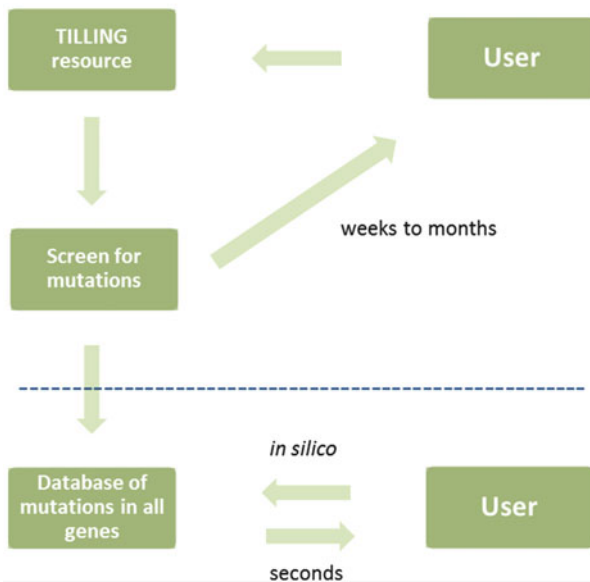


Fig. 3 Traditional TILLING services versus in silico resources. In traditional TILLING services screening for mutations begins when a user requests mutations in a specific gene region (top). Screening of the population is performed for that target and identified mutations are reported back to the user along with information on how to access seed stock. Depending on the speed of the TILLING facility and the number of orders placed, it may take weeks to months before the user receives results [21]. In in silico TILLING, all mutations in a population are discovered and catalogued in a database prior to any user requests. The user searches a database for mutations in the selected gene and results are provided in the time it takes for the search to be completed, typically seconds

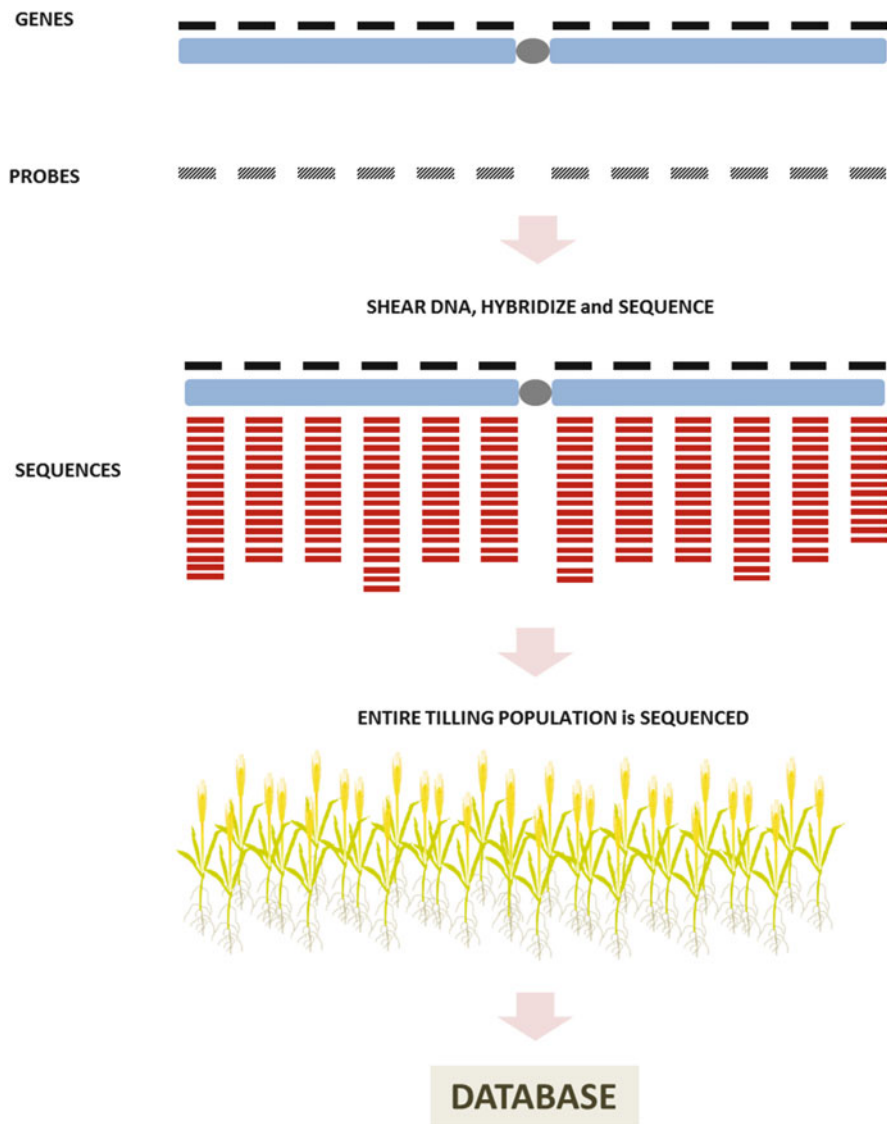


Fig. 4 A simplified example of developing an in silico TILLING resource using reduced representation exome capture sequencing. Probes are designed to cover all genomic sequences of interest (top). Exon sequences that code for proteins are good targets for mutagenesis as the effect of mutations can be predicted in advance. Probes can be designed for any region such as promoters and other regulatory elements. Genomic DNA from mutant plants is isolated, sheared, and then hybridized to the probes. DNA-Probe hybrids are then physically separated and sequenced (middle). This process is performed on the entire TILLING population and a database of mutation information for each plant is created (bottom). Users of the resource access the database and search for mutations in their specific gene target(s). A list of plants harboring identified mutations is returned along with information on how to access seed for the selected plants

covered approximately 60 Mbp of exonic sequence. The size of the zebrafish genome is ~1.4 Gb and so the exome approach represents a major reduction in sequencing loads (about 23 \times) while not reducing the ability to identify genic mutations.

The approach is especially appealing for large genome plants where there has already been reported success in TILLING. There have been many examples of successful TILLING in polyploid wheat with high mutation densities [18–20, 75]. Reverse-genetics is a powerful approach in polyploids where recessive mutations are not observed due to the presence of homeologous sequences that must also be mutated before a phenotype can be observed. Slade and colleagues combined mutations in starch branching IIa genes in the A, B, and D genomes to produce high amylose wheat [75]. Uauy and colleagues have taken a similar approach of combining mutations in the genomes to increase grain size [76]. An *in silico* TILLING resource has been produced for both tetraploid and hexaploid wheat. More than 10 million mutations have been reported, making it the largest dataset on the effects of EMS mutagenesis on a plant genome [62]. It is likely that the success of this project will stimulate similar endeavors in other important plants.

5 Reverse-Genetics Using Targeted Genome Editing

Huge progress has been made in targeted genome editing within the past few years [77]. A pubmed search for the term “CRISPR” performed 20 February 2018 showed a total of 8,340 hits with a 30% increase between 2016 and 2017. It is a safe estimation that this number will be much higher and there will be many new breakthroughs by the time this book chapter is published. Targeted genome editing, as the name implies, involves the generation of a genomic change of a precise type in a precise location in the genome of a plant, animal, or microorganism. It is thus a reverse-genetic technique that utilizes induced mutations and therefore shares many similarities to TILLING. With the exception of relatively rare off-target mutations, the approach has the advantage that only the desired change is produced in the organism. A variety of methods and variations on methods have been described including Meganucleases, Zinc finger nucleases (ZFNs), Transcription Activator-Like Effector-based Nucleases (TALENs), and RNA-guided editing using the CRISPR/Cas system [78]. The nucleases create double-strand breaks (DSBs) at desired sequence-specific locations in the genome, following which the DSBs sites are repaired either by non-homologous end joining (NHEJ) or homology-directed repair (HDR) mechanisms that result in the fixation of mutations in the genomic sequence.

Differences between the above-mentioned engineered nucleases have been extensively reviewed [78]. The simplicity of the CRISPR/Cas system has enabled it to become the predominate genome editing method. It is based on the bacterial CRISPR/Cas type II prokaryotic adaptive immune system and uses a Cas9 nuclease and only one engineered single-guide RNA (sgRNA) to specify the target DNA

sequence. In addition to creating novel specific sequence changes, there is an added advantage that homozygous mutations can be immediately produced in a single generation [79, 80]. Further, homeologous loci in polyploid species can be simultaneously edited as was shown by Qiu and colleagues in their work procuring resistance to powdery mildew in hexaploid wheat by mutating three *MLO* loci [81]. Modifications such as CRISPR interference (CRISPRi) and CRISPR activator (CRISPRa) allow modulation of gene expression that can be used for plant studies including pathway analysis of plant stress response [82]. While the focus of this chapter is on plant sciences, it should be noted that the CRISPR based approaches hold tremendous potential to revolutionize human health through the development of disease models and the direct correction of deleterious (disease causing) variants in human cells [83]. Modification of human embryos has been described, something that was merely a trope of science fiction a scant decade ago [84, 85]. The ethical and regulatory issues of using CRISPR approaches in humans, as well as regulatory and social acceptance issues of their use in crops are still being promulgated.

6 Choosing the Best Approach

Given the choices of forward- versus reverse-genetics and random versus targeted mutagenesis, one can consider the comparative advantages of the different approaches to meet breeding and research objectives (Table 2). For example, forward-genetics has been a mainstay of basic research and breeding for decades. Advantages include the fact that it is phenotype driven and no prior knowledge of gene function is required for success. Indeed, the first mutant crop variety was released in the 1930s long before DNA was shown to be the genetic material. There is no intellectual property or regulation when using induced mutations in crop breeding programs and it can be initiated cheaply and easily in any country, including developing ones. This may be one reason why mutation breeding has been so successful and resulted in the addition of billions of dollars to economies [5, 14].

Table 2 A comparison of forward- and reverse-genetics and random versus targeted mutagenesis

	Random mutagenesis and phenotyping	TILLING	CRISPR/Cas
Method type	Forward-genetics	Reverse-genetics	Reverse-genetics
Knowledge of genes/alleles required?	No/No	Yes/No	Yes/Yes
Procedure for inducing variation	Random mutagenesis	Random mutagenesis	Targeted mutagenesis
Target specificity?	No	No	Yes
Regulated?	No	No	No policy yet in some countries
Issues	Possible genetic linkage of induced mutations	Possible genetic linkage of induced mutations	Off target events

Reverse genetics by TILLING requires knowledge of candidate gene targets, but not of specific alleles. An advantage with TILLING is that populations can be prepared in advance where allelic series are available in all genes so that both knockout and missense changes can be recovered by researchers as quickly as seed can be sent from a stock center. Multiple alleles can be tested directly to deepen knowledge on gene function. With advances such as exome capture sequencing, the development of in silico TILLING resources will become inexpensive and common. The major disadvantage with TILLING is the fact that any plant may harbor thousands of point mutations and several backcrosses may be required to unambiguously assign gene function. This is relatively straightforward in genetically tractable crops like cereals but can become extremely challenging in crops like triploid bananas, which are obligate vegetatively propagated. With targeted genome editing approaches one must design and create each mutation. This is considerably more up-front work than random chemical mutagenesis. However, one can avoid the issue of background mutations/linkage drag and make a “clean” variant. Further, the ability to make homozygous lesions has great potential in obligate vegetatively propagated crops like triploid banana, where creating and utilizing recessive alleles is laborious.

When considering forward- versus reverse-genetics and random versus targeted mutagenesis, it is likely that many researchers will not treat these as either/or propositions but rather choose a combinatorial approach that allows the quickest and most cost-effective means to reach his or her goal. One can imagine, for example, using an in silico TILLING resource to first test and validate gene function and then later using CRISPR to create a single mutation in an elite breeding cultivar. This could in some cases be substantially faster than traditional introgression and would avoid any problems with genetically-linked induced mutations that might be present in TILLING lines. The opposite approach could also be taken if targeted genome editing is not desired in the final product. Once genes and alleles are validated by CRISPR, a traditional TILLING population could be created to generate the desired improved trait. Forward-genetics will remain powerful for gene discovery and new sequencing based approaches to cloning mutant alleles will provide information on genes and variants causative for phenotypes that can support reverse approaches [86, 87].

7 Concluding Remarks and Future Perspectives

Genetic mutations and recombination allowed the evolution of species, domestication of plants and animals, and provides the diversity required for modern plant breeding. New technological developments have meant that mutations remain a fundamental tool for both breeder and basic researcher. The advent of reverse-genetics in the 1980s marked the beginning of a new way to use mutations through disruption of specific genes of interest. The vast amount of gene sequence available means that reverse-genetics can be considered for many species. Indeed, whole genome sequences are now available for 47 important crops [88]. TILLING is easily

adapted for most crops as it relies on traditional chemical mutagenesis. A variety of mutation discovery methods can be efficiently used in TILLING screens and so it is expected that TILLING will remain an important approach for functional genomics studies and for breeding. In silico TILLING has been established in wheat, one of the most important food crops. As mutation discovery technologies improve and large-scale sequencing becomes cheap and commonplace, it is expected that in silico TILLING resources will become standard for many plant research communities. Targeted genome editing complements random mutagenesis. As methods such as CRISPR/Cas become routine, the genetic toolkit for many plant species will expand further. This will allow fundamental new biological insights, and also the improvement and domestication of plants that have great potential to help address growing pressures on global food security.

Acknowledgements Funding for the authors was provided by the FAO/IAEA Joint Programme.

References

1. De Vries H (1901) Die mutationstheorie. Veit & Co, Leipzig
2. Nei M, Nozawa M (2011) Roles of mutation and selection in speciation: from Hugo de Vries to the modern genomic era. *Genome Biol Evol* 3:812–829
3. Crow EW, Crow JF (2002) 100 years ago: Walter Sutton and the chromosome theory of heredity. *Genetics* 160:1–4
4. Shu QY, Forster BP, Nakagawa H (2012) Plant mutation breeding and biotechnology. CABI International, Cambridge
5. Ahloowalia BS, Maluszynski M, Nichterlein K (2004) Global impact of mutation-derived varieties. *Euphytica* 135:187–204
6. Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92–94
7. Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen JQ, Hurst LD, Tian D (2015) Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523:463–467
8. Muller HJ (1927) Artificial transmutation of the gene. *Science* 66:84–87
9. Stadler LJ (1928) Genetic effects of X-rays in maize. *Proc Natl Acad Sci U S A* 14:69–75
10. Stadler LJ (1928) Mutations in barley induced by X-rays and radium. *Science* 68:186–187
11. Tollenaar D (1934) Untersuchungen ueber Mutation bei Tabak. I. Entstehungsweise und Wesen kuenstlich erzeugter Gen-Mutanten. *Genetica* 16:111–152
12. Tollenaar D (1938) Untersuchungen ueber Mutation bei Tabak. II. Einige kuenstlich erzeugte Chromosom-Mutanten. *Genetica* 20:285–294
13. IAEA (2016) Mutant variety database [Online]. Available: <http://mvd.iaea.org/>. Accessed 2 May 2016
14. Jankowicz-Cieslak J, Till BJ (2015) Forward and reverse genetics in crop breeding. In: Al-Khayri JM et al (eds) *Advances in plant breeding strategies: breeding, biotechnology and molecular tools*. Springer, Berlin
15. Hutchison CA (2007) DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* 35:6227–6237
16. Jankowicz-Cieslak J, Huynh OA, Brozynska M, Nakitandwe J, Till BJ (2012) Induction, rapid fixation and retention of mutations in vegetatively propagated banana. *Plant Biotechnol J* 10:1056–1066

17. Till BJ, Reynolds SH, Weil C, Springer N, Burtner C, Young K, Bowers E, Codomo CA, Enns LC, Odden AR, Greene EA, Comai L, Henikoff S (2004) Discovery of induced point mutations in maize genes by TILLING. *BMC Plant Biol* 4:12
18. Dong C, Dalton-Morgan J, Vincent K, Sharp P (2009) A modified TILLING method for wheat breeding. *Plant Genome* 2:39–47
19. Slade AJ, Fuerstenberg SI, Loeffler D, Steine MN, Facciotti D (2005) A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. *Nat Biotechnol* 23:75–81
20. Uauy C, Paraiso F, Colasuonno P, Tran RK, Tsai H, Berardi S, Comai L, Dubcovsky J (2009) A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. *BMC Plant Biol* 9:115
21. Till BJ, Reynolds SH, Greene EA, Codomo CA, Enns LC, Johnson JE, Burtner C, Odden AR, Young K, Taylor NE, Henikoff JG, Comai L, Henikoff S (2003) Large-scale discovery of induced point mutations with high-throughput TILLING. *Genome Res* 13:524–530
22. Auerbach C, Robson JM, Carr JG (1947) The chemical production of mutations. *Science* 105:243–247
23. Coen ES, Meyerowitz EM (1991) The war of the whorls: genetic interactions controlling flower development. *Nature* 353:31–37
24. Hartwell LH, Culotti J, Reid B (1970) Genetic control of the cell-division cycle in yeast. I. Detection of mutants. *Proc Natl Acad Sci U S A* 66:352–359
25. Nusslein-Volhard C, Wieschaus E (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287:795–801
26. Haughn GW, Somerville CR (1987) Selection for herbicide resistance at the whole plant level. In: Lebaron HM, Mumma RO, Hoenycutt RC, Duesing JH (eds) *Applications of biotechnology to agricultural chemistry*. American Chemical Society, Washington, DC
27. Koornneef M, Jorna ML, Brinkhorst-Van Der Swan DL, Karszen CM (1982) The isolation of abscisic acid (ABA) deficient mutants by selection of induced revertants in non-germinating gibberellin sensitive lines of *Arabidopsis thaliana* (L.) Heynh. *Theor Appl Genet* 61:385–393
28. Bentley A, MacLennan B, Calvo J, Dearolf CR (2000) Targeted recovery of mutations in *Drosophila*. *Genetics* 156:1169–1173
29. McCallum CM, Comai L, Greene EA, Henikoff S (2000) Targeted screening for induced mutations. *Nat Biotechnol* 18:455–457
30. Kurowska M, Daszkowska-Golec A, Gruszka D, Marzec M, Szurman M, Szarejko I, Maluszynski M (2011) TILLING – a shortcut in functional genomics. *J Appl Genet* 52:371–390
31. Dodgson JB, Wells RD (1977) Action of single-strand specific nucleases on model DNA heteroduplexes of defined size and sequence. *Biochemistry* 16:2374–2379
32. Shenk TE, Rhodes C, Rigby PW, Berg P (1975) Biochemical method for mapping mutational alterations in DNA with S1 nuclease: the location of deletions and temperature-sensitive mutations in simian virus 40. *Proc Natl Acad Sci U S A* 72:989–993
33. Silber JR, Loeb LA (1981) S1 nuclease does not cleave DNA at single-base mis-matches. *Biochim Biophys Acta* 656:256–264
34. Till BJ, Burtner C, Comai L, Henikoff S (2004) Mismatch cleavage by single-strand specific nucleases. *Nucleic Acids Res* 32:2632–2641
35. Colbert T, Till BJ, Tompa R, Reynolds S, Steine MN, Yeung AT, McCallum CM, Comai L, Henikoff S (2001) High-throughput screening for induced point mutations. *Plant Physiol* 126:480–484
36. Jankowicz-Cieslak J, Huynh OA, Bado S, Matijevic M, Till BJ (2011) Reverse-genetics by TILLING expands through the plant kingdom. *Emir J Food Agric* 23:290–300
37. ATP (2007) History of ATP [Online]. Available: http://tilling.fhcr.org/files/History_of_ATP.html. Accessed 2 May 2016
38. Cooper JL, Till BJ, Laport RG, Darlow MC, Kleffner JM, Jamai A, El-Mellouki T, Liu S, Ritchie R, Nielsen N, Bilyeu KD, Meksem K, Comai L, Henikoff S (2008) TILLING to detect induced mutations in soybean. *BMC Plant Biol* 8:9

39. Till BJ, Cooper J, Tai TH, Colowit P, Greene EA, Henikoff S, Comai L (2007) Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biol* 7:19
40. Greene EA, Codomo CA, Taylor NE, Henikoff JG, Till BJ, Reynolds SH, Enns LC, Burtner C, Johnson JE, Odden AR, Comai L, Henikoff S (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. *Genetics* 164:731–740
41. Tsai H, Missirian V, Ngo KJ, Tran RK, Chan SR, Sundaresan V, Comai L (2013) Production of a high-efficiency TILLING population through polyploidization. *Plant Physiol* 161:1604–1614
42. Knoll JE, Ramos ML, Zeng YJ, Holbrook CC, Chow M, Chen SX, Maleki S, Bhattacharya A, Ozias-Akins P (2011) TILLING for allergen reduction and improvement of quality traits in peanut (*Arachis hypogaea* L.) *BMC Plant Biol* 11:81
43. Guo Y, Abernathy B, Zeng Y, Ozias-Akins P (2015) TILLING by sequencing to identify induced mutations in stress resistance genes of peanut (*Arachis hypogaea*). *BMC Genomics* 16:157
44. Gilchrist EJ, Sidebottom CH, Koh CS, Macinnes T, Sharpe AG, Haughn GW (2013) A mutant *Brassica napus* (canola) population for the identification of new genetic diversity via TILLING and next generation sequencing. *PLoS One* 8:e84303
45. Zhu Q, Smith SM, Ayele M, Yang L, Jogi A, Chaluvadi SR, Bennetzen JL (2012) High-throughput discovery of mutations in *tef* semi-dwarfing genes by next-generation sequencing analysis. *Genetics* 192:819–829
46. Sabetta W, Alba V, Blanco A, Montemurro C (2011) SunTILL: a TILLING resource for gene function analysis in sunflower. *Plant Methods* 7:20
47. Kumar AP, Boualem A, Bhattacharya A, Parikh S, Desai N, Zambelli A, Leon A, Chatterjee M, Bendahmane A (2013) SMART—sunflower mutant population and reverse genetic tool for crop improvement. *BMC Plant Biol* 13:38
48. Caldwell DG, McCallum N, Shaw P, Muehlbauer GJ, Marshall DF, Waugh R (2004) A structured mutant population for forward and reverse genetics in Barley (*Hordeum vulgare* L.) *Plant J* 40:143–150
49. Gottwald S, Bauer P, Komatsuda T, Lundqvist U, Stein N (2009) TILLING in the two-rowed barley cultivar ‘Barke’ reveals preferred sites of functional diversity in the gene *HvHox1*. *BMC Res Notes* 2:258
50. Rigola D, Van Oeveren J, Janssen A, Bonne A, Schneiders H, Van Der Poel HJ, Van Orsouw NJ, Hogers RC, De Both MT, Van Eijk MJ (2009) High-throughput detection of induced mutations and natural variation using keypoint technology. *PLoS One* 4:e4761
51. Tsai H, Howell T, Nitcher R, Missirian V, Watson B, Ngo KJ, Lieberman M, Fass J, Uauy C, Tran RK, Khan AA, Filkov V, Tai TH, Dubcovsky J, Comai L (2011) Discovery of rare mutations in populations: TILLING by sequencing. *Plant Physiol* 156:1257–1268
52. Suzuki T, Eiguchi M, Kumamaru T, Satoh H, Matsusaka H, Moriguchi K, Nagato Y, Kurata N (2008) MNU-induced mutant pools and high performance TILLING enable finding of any gene mutation in rice. *Mol Gen Genomics* 279:213–223
53. Chen L, Huang L, Min D, Phillips A, Wang S, Madgwick PJ, Parry MA, Hu YG (2012) Development and characterization of a new tilling population of common bread wheat (*Triticum aestivum* L.) *PLoS One* 7:e41570
54. Rawat N, Sehgal SK, Joshi A, Rothe N, Wilson DL, McGraw N, Vadlani PV, Li W, Gill BS (2012) A diploid wheat TILLING resource for wheat functional genomics. *BMC Plant Biol* 12:205
55. REVGENUK (2016) Revgen UK search databases [Online]. Available: <http://revgenuk.jic.ac.uk/search-databases/>. Accessed 2 May 2016
56. UC-Davis TC (2016) TILLING at the UC Davis TILLING core [Online]. Available: http://tilling.ucdavis.edu/index.php/Main_Page. Accessed 2 May 2016
57. URGV_TILLING (2016) UTiLLdb URGV TILLING database [Online]. Available: <http://urgv.evry.inra.fr/UTiLLdb>. Accessed 2 May 2016
58. Welcome_Trust_Sanger_Institute (2016) Zebrafish mutation project [Online]. Available: <http://www.sanger.ac.uk/resources/zebrafish/zmp/>. Accessed 2 May 2016

59. Wheat_TILLING (2016) Wheat TILLING at the John Innes Centre [Online]. Available: <http://www.wheat-tiling.com/>. Accessed 2 May 2016
60. Segal GA (1984) A review of the genetic effects of ethyl methanesulfonate. *Mutat Res* 134:113–142
61. Henry IM, Nagalakshmi U, Lieberman MC et al (2014) Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *Plant Cell*. <https://doi.org/10.1105/tpc.113.121590>
62. Krasileva KV, Vasquez-Gross HA, Howell T et al (2017) Uncovering hidden variation in polyploid wheat. *Proc Natl Acad Sci U S A* 114:E913–E921
63. Comai L (2016) ComaiWIKI TILLING by sequencing [Online]. Available: http://comailab.genomecenter.ucdavis.edu/index.php/TILLING_by_Sequencing#Bioinformatics_tools. Accessed 2 May 2016
64. Chi X, Zhang Y, Xue Z, Feng L, Liu H, Wang F, Qi X (2014) Discovery of rare mutations in extensively pooled DNA samples using multiple target enrichment. *Plant Biotechnol J* 12:709–717
65. Pan L, Shah AN, Phelps IG, Doherty D, Johnson EA, Moens CB (2015) Rapid identification and recovery of ENU-induced mutations with next-generation sequencing and paired-end low-error analysis. *BMC Genomics* 16:83
66. Gupta P, Reddaiah B, Salava H et al (2017) Next-generation sequencing (NGS)-based identification of induced mutations in a doubly mutagenized tomato (*Solanum lycopersicum*) population. *Plant J* 92:495–508
67. Mccallum CM, Comai L, Greene EA, Henikoff S (2000) Choosing optimal regions for TILLING. *Plant Physiol* 123:439–442
68. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814
69. Taylor NE, Greene EA (2003) PARSESNP: a tool for the analysis of nucleotide polymorphisms. *Nucleic Acids Res* 31:3808–3811
70. Hudson K, Thapa R, Rainey KM (2016) PAG XXIV W601 identification of rare alleles in soybean using TILLING by sequencing [Online]. Available: <https://pag.confex.com/pag/xxiv/webprogram/Paper20560.html>. Accessed 2 May 2016
71. Duitama J, Kafuri L, Tello D et al (2017) Deep assessment of genomic diversity in cassava for herbicide tolerance and starch biosynthesis. *Comput Struct Biotechnol J* 15:185–194. <https://doi.org/10.1016/j.csbj.2017.01.002>
72. Guo L, Gao Z, Qian Q (2014) Application of resequencing to rice genomics, functional genomics and evolutionary analysis. *Rice (N Y)* 7:4
73. Weigel D, Mott R (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* 10:107
74. Kettleborough RN, Busch-Nentwich EM, Harvey SA, Dooley CM, De Bruijn E, Van Eeden F, Sealy I, White RJ, Herd C, Nijman IJ, Fenyves F, Mehroke S, Scahill C, Gibbons R, Wali N, Carruthers S, Hall A, Yen J, Cuppen E, Stemple DL (2013) A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* 496:494–497
75. Slade AJ, Mcguire C, Loeffler D, Mullenberg J, Skinner W, Fazio G, Holm A, Brandt KM, Steine MN, Goodstal JF, Knauf VC (2012) Development of high amylose wheat through TILLING. *BMC Plant Biol* 12:69
76. Simmonds J, Scott P, Brinton J et al (2016) A splice acceptor site mutation in TaGW2-A1 increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. *Theor Appl Genet* 129:1099–1112
77. Mei Y, Wang Y, Chen H, Sun ZS, Ju XD (2016) Recent progress in CRISPR/Cas9 technology. *J Genet Genomics* 43:63–75
78. Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol* 31:397–405

79. Brooks C, Nekrasov V, Lippman ZB, Van Eck J (2014) Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. *Plant Physiol* 166:1292–1297
80. Zhang H, Zhang J, Wei P, Zhang B, Gou F, Feng Z, Mao Y, Yang L, Zhang H, Xu N, Zhu JK (2014) The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant Biotechnol J* 12:797–807
81. Wang Y, Cheng X, Shan Q, Zhang Y, Liu J, Gao C, Qiu JL (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nat Biotechnol* 32:947–951
82. Khatodia S, Bhatotia K, Passricha N, Khurana SMP, Tuteja N (2016) The CRISPR/Cas genome-editing tool: application in improvement of crops. *Front Plant Sci* 7:506
83. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533:420–424
84. Kang X, He W, Huang Y, Yu Q, Chen Y, Gao X, Sun X, Fan Y (2016) Introducing precise genetic modifications into human 3PN embryos by CRISPR/Cas-mediated genome editing. *J Assist Reprod Genet* 33:581–588
85. Liang P, Xu Y, Zhang X, Ding C, Huang R, Zhang Z, Lv J, Xie X, Chen Y, Li Y, Sun Y, Bai Y, Songyang Z, Ma W, Zhou C, Huang J (2015) CRISPR/Cas9-mediated gene editing in human triprenuclear zygotes. *Protein Cell* 6:363–372
86. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, Matsumura H, Mitsuoka C, Tamiru M, Innan H, Cano L, Kamoun S, Terauchi R (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat Biotechnol* 30:174–178
87. Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Methods* 6:550–551
88. Huang S, Weigel D, Beachy RN, Li J (2016) A proposed regulatory framework for genome-edited crops. *Nat Genet* 48:109–111

Advances in Transcriptomics of Plants



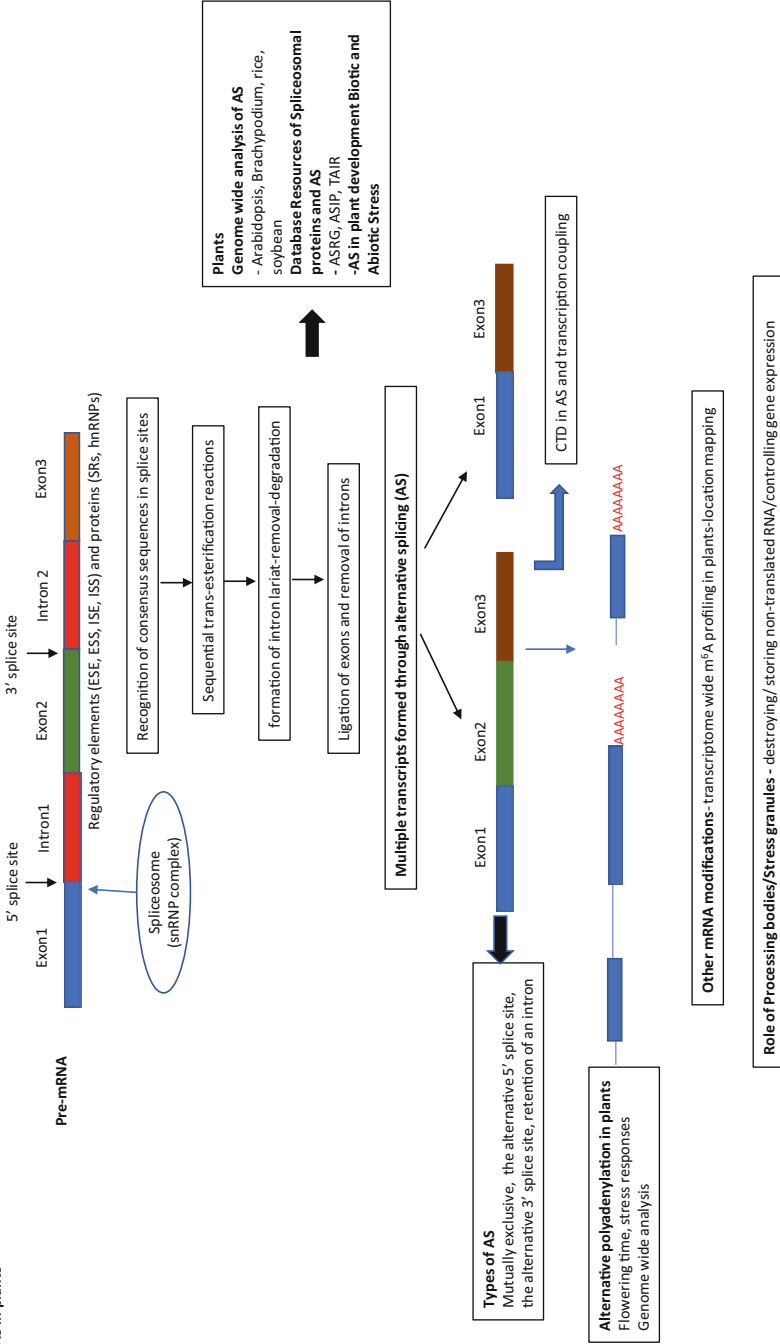
Naghmeh Nejat, Abirami Ramalingam, and Nitin Mantri

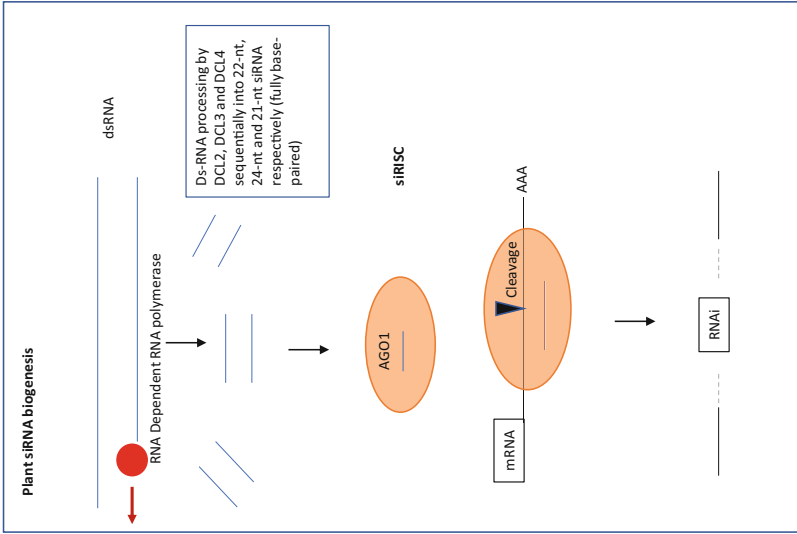
Abstract The current global population of 7.3 billion is estimated to reach 9.7 billion in the year 2050. Rapid population growth is driving up global food demand. Additionally, global climate change, environmental degradation, drought, emerging diseases, and salty soils are the current threats to global food security. In order to mitigate the adverse effects of these diverse agricultural productivity constraints and enhance crop yield and stress-tolerance in plants, we need to go beyond traditional and molecular plant breeding. The powerful new tools for genome editing, Transcription Activator-Like Effector Nucleases (TALENs) and Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR)/Cas systems (CRISPR-Cas9), have been hailed as a quantum leap forward in the development of stress-resistant plants. Plant breeding techniques, however, have several drawbacks. Hence, identification of transcriptional regulatory elements and deciphering mechanisms underlying transcriptional regulation are crucial to avoiding unintended consequences in modified crop plants, which could ultimately have negative impacts on human health. RNA splicing as an essential regulated post-transcriptional process, alternative polyadenylation as an RNA-processing mechanism, along with non-coding RNAs (microRNAs, small interfering RNAs and long non-coding RNAs) have been identified as major players in gene regulation. In this chapter, we highlight new findings on the essential roles of alternative splicing and alternative polyadenylation in plant development and response to biotic and abiotic stresses. We also discuss biogenesis and the functions of microRNAs (miRNAs) and small interfering RNAs (siRNAs) in plants and recent advances in our knowledge of the roles of miRNAs and siRNAs in plant stress response.

N. Nejat, A. Ramalingam, and N. Mantri (✉)
The Pangenomics Group, School of Science, RMIT University, Melbourne, VIC, Australia
e-mail: nitin.mantri@rmit.edu.au

Graphical Abstract

AS in plants





miRNA function in plants

- Analysis in Arabidopsis, tomato, oilseed rape
- Cell signalling, differentiation, DNA damage repair,
- hormone signalling, heterosis

Biotic stress responses

- against bacterial, fungal and viral pathogens
- miR156, miR159, miR172, miR319, miRR393

Abiotic stress responses

- involving cold, heat, heavy metals nutrient, oxidation,
- UV-B,
- miR319, miR169

siRNA function in plants

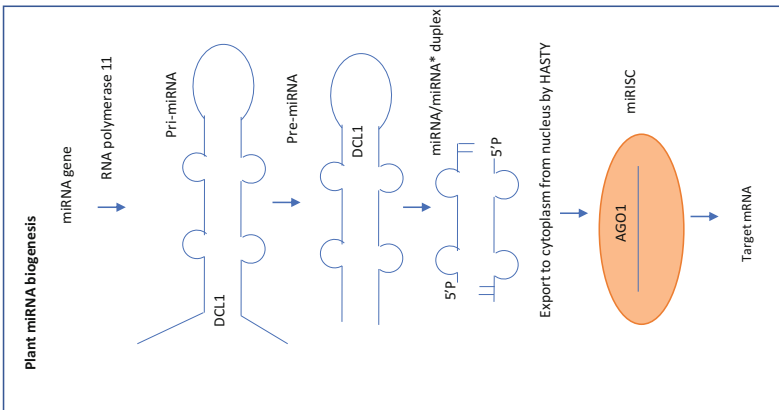
- hp-siRNAs, nat-siRNAs, het-siRNA, tasiRNA, phasiRNA, easiRNA
- discovered in plants with deep sequencing
- development, hormone signalling, fertility & reproduction, biotic and abiotic stress

Biotic stress responses

- nat-siRNA, against bacterial pathogens

Abiotic stress responses

- nat-siRNA, in response to cold, drought, salt stress.
- Arabidopsis, wheat



Keywords Abiotic stress, Biotic stress, Alternative splicing, Alternative polyadenylation, microRNAs, Small interfering RNAs

Contents

1	Introduction	164
2	Alternative Splicing, Alternative Polyadenylation, and Other Modifications of mRNA ..	166
2.1	Types of Alternative Splicing (AS)	168
2.2	Coupling of Transcription with AS	168
2.3	AS in Plants	168
2.4	Database Resources of Plant Spliceosomal Proteins and AS	169
2.5	Role in Plant Development	169
2.6	Role in Biotic and Abiotic Stress Response	169
2.7	Alternative Polyadenylation	170
2.8	Modifications in mRNA	170
2.9	Stress Response Mechanism and the Cytoplasmic RNA-Containing Granules	172
3	microRNAs (miRNAs) and Small Interfering RNAs (siRNAs)	172
3.1	Biogenesis of miRNAs in Plants	173
3.2	Biogenesis of siRNAs in Plants	173
3.3	Functions of miRNAs and siRNAs in Plants	174
3.4	Role of miRNAs in Plant Stress Responses	174
3.5	miRNAs in Biotic Stress	175
3.6	miRNAs in Abiotic Stress	176
3.7	Role of siRNAs in Plant Stress Responses	177
3.8	siRNAs in Biotic Stress	177
3.9	siRNAs in Abiotic Stress	178
4	Conclusions and Future Prospects	178
	References	179

1 Introduction

Biotic and abiotic stresses, global climate change, and environmental pollution have a significant negative impact on crop yields. Together with rapid global population growth, these factors threaten global food security. The current world population of 7.3 billion is expected to reach 8.5 billion by 2030, 9.7 billion in 2050 and 11.2 billion in 2100, according to the most recent UN DESA report, “World Population Prospects: The 2015 Revision” [1]. Moreover, food demand is expected to increase by 59–98% between 2005 and 2050 [2]. In order to mitigate and control these diverse agricultural productivity constraints, extensive effort has been put into improving crop yield and stress-tolerance through traditional and molecular plant breeding. Traditional or conventional plant breeding is a time-consuming and labor-intensive approach. It is limited to the exchange of genes between fairly closely related species [3]. In order to overcome these hurdles, molecular breeding through gene manipulation has widely been used to develop new high-yielding, stress-tolerant crop varieties [4]. However, this technology has some limitations such as non-targeted and unanticipated effects [5]. Recently, new technologies such as Transcription Activator-Like Effector Nucleases (TALENs) and Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR)/Cas systems have emerged

for genome editing [6]. However, harnessing the benefits of these technologies requires a complete understanding of the complexity of plant defense mechanisms and stress signaling pathways at the molecular level. This knowledge is crucial to enable development of high-yielding, stress-resistant crop plants with minimum yield penalty through selective genetic engineering and precise gene editing techniques [7].

Over the past 20 years, the field of gene expression profiling has undergone a dramatic revolution. Transcriptomics has witnessed remarkable success due to major advances in transcriptome sequencing and analysis technologies. A wide variety of molecular biology techniques have been used for expression profiling and transcription quantification. Traditional techniques such as northern blotting and in situ hybridization [8] and reverse transcription polymerase chain reaction (RT-PCR) [9] allow only single transcripts or small groups of transcripts to be analyzed at once. The real-time RT-PCR method is a medium-throughput and very sensitive technique for the detection of low-abundance mRNA. It has been widely used for absolute and relative gene expression quantification [10–12]. The development of microarrays in the mid-1990s revolutionized gene expression studies and provided a new tool for genome expression profiling by allowing large-scale analysis of thousands of genes simultaneously [13]. Microarrays have widely been employed to understand molecular mechanisms underlying plant development and response to a multitude of stresses [14–17]. More recently, next-generation sequencing (NGS) has essentially provided the second revolution since the development of microarrays. Through high-throughput sequencing, it has remarkably improved our understanding and knowledge of gene regulatory mechanisms and epigenetics [18, 19]. NGS-based RNA sequencing (RNA-seq) allows detection and quantification of known, novel and rare transcripts, genome annotation, and rearrangement detection to non-coding RNA discovery. Furthermore, it provides greater insights into biological pathways and molecular mechanisms that regulate cell fate, development, and disease progression [6, 18, 20].

Recent advances in transcriptomics technologies shed light on the dark intergenic regions between protein-coding genes traditionally referred to as transcriptional “noise,” “junk DNA,” or experimental artifact. In 2012, ENCODE (Encyclopedia of DNA Elements) declared that 80% of the human genome has a biochemical function. However, scientists of the ENCODE project recently got together in Potomac, MD, USA and claimed that ~50% is functional [21]. By contrast, Rands and Colleagues claim that 8.2% of the human genome is likely to be functional [22]. In comparison, only a tiny portion of the transcribed human genome (~1–2%) codes for proteins [23]. The number of protein-coding genes in the human genome is reported to be fewer than 20,000 genes and has continued to shrink [24]. According to the most recent estimate of the GENCODE annotation of the human genome, GENCODE release 24 (09.12.2015) corresponds to Ensembl 83, 84, and the human genome encompasses 19,815 protein-coding genes and 25,823 non-coding RNA genes (ncRNAs). Of these, 15,941 and 9,882 are long and small non-coding RNA genes, respectively. Moreover, many new non-coding RNA classes have been identified in the last few years and classified based on their distinct biogenesis pathways. Although little is known about plant genomics and plant genome

composition, several recent studies have identified different types of non-coding RNAs with diverse functions in plants [25]. In this chapter, we discuss recent advances in our knowledge of the biological functions of mRNAs (with a particular focus on alternatively spliced mRNAs and polyadenylation) and ncRNAs (with a particular focus on miRNAs and siRNAs) in plant development and response to biotic and abiotic stresses. The purpose of this chapter is to provide an overview of important regulatory components, apart from pure mRNA expression, which has been studied for several decades.

2 Alternative Splicing, Alternative Polyadenylation, and Other Modifications of mRNA

The pre-mRNA containing introns can be alternatively spliced to generate multiple transcripts from a single gene through the differential use of splice sites that increase the transcriptome and proteome complexity of the cells and tissues [26, 27]. Eighty percent of the genes in plants and animals contain introns. Splicing, which is the removal of introns, is carried out by the spliceosome that surrounds the splice sites at each intron. The pre-mRNA (primary transcript) structure includes *cis*-elements such as the 5' splice site, the branch-point that is close to the 3' end splice site, the polypyrimidine ring tract, and the 3' splice site, which are required for splicing. The splice sites contain consensus sequences that are recognized by the spliceosome. The spliceosome, which is a complex ribonucleoprotein mega particle, consists of small nuclear ribonucleoproteins (snRNPs)- U1, U2, U4, U5, U6, and auxiliary factors, U2AF65 and U2AF35. Spliceosomes have a stronger affinity for some splice sites and a weaker affinity for others, and this phenomenon is important in alternative splicing (AS) [28]. The spliceosome recognizes these features and it applies two sequential *trans*-esterification reactions that ligate the selected exon sequences and remove the introns. The first step involves the nucleophilic attack by the 2'OH group of an important adenosine in the branch consensus site on the 5' splice site, forming a branched RNA intermediate called intron lariat. After this, some of the snRNPs are released. In the second step, the 3'OH group of the upstream exon targets the 3' splice site. This reaction results in the spliced mRNA and the intron lariat is removed and degraded [28]. The *cis*-regulatory sequences in the pre-mRNA (exon splicing enhancers (ESE), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs)) as well as the AS regulatory proteins (Ser/Arg-rich proteins (SRs), and heterogeneous nuclear ribonucleoproteins (hnRNPs)) are important regulators of the splicing process [29, 30]. The exon-exon junction complex (EJC) accumulates 24 nt upstream of the exon-exon junction for exportation of mRNA from the nucleus and for the cytoplasmic mRNA control [31]. The binding of the serine/arginine-rich (SR) family of splicing factors to ESE helps in the recruitment of the splicing components and prevents "exon skipping" [32]. The key studies on mRNA modification are summarized in Table 1.

Table 1 Key studies on mRNA modification, and miRNA and siRNA biogenesis

Features	Steps in biogenesis	References
<i>mRNA modification</i>		
AS	Recognition of consensus sequences in splice sites by the spliceosome	Kornblihtt et al. [28]
	Application of <i>trans</i> -esterification reactions by spliceosome which ligates selected exons and removes introns	
	Cis-regulatory sequences in the pre-mRNA are important regulators of the splicing process	Wang et al. [29] and Wang and Burge [30]
	EJC accumulation 2 nt upstream of the exon-exon junction for exportation of mRNA and cytoplasmic mRNA control	Le Hir and Anderson [31]
Forms of AS	Sequences of exons and introns are included or excluded from the mRNA based on AS (Formation of cassette exon, mutually exclusive exons, etc.)	Reddy [27]
Coupling of transcription with AS	AS is co-transcriptional, in which the CTD is involved	Dujardin et al. [33]
Alternative polyadenylation	Production of different transcripts with altered coding capacity	Xing and Li [34]
Modifications in mRNA	Involvement of N ⁶ -methyladenosine (m ⁶ A), 5-methylcytosine (m ⁵ C) and pseudouridine (ψ)	Li et al. [35], Gilbert et al. [36] and Shen et al. [37]
<i>miRNAs and siRNAs</i>		
Processing of miRNA	Formation of precursor miRNA involving DCL1	Bartel [38] and Bologna and Vionnet [39]
	Involvement of DCL1 in the formation of mature miRNA	Jones-Rhoades et al. [40]
	Exportation of miRNA from the nucleus	Bollman et al. [41] and Park et al. [42]
	The miRISC is loaded onto an Argonaute protein family member and is guided to the targeted mRNA	Bartel [43] and Meister [44]
Processing of siRNA	Cleaving of long duplex RNA structures by DCL3 and DCL4 into 22-nt, 24-nt, and 21-nt siRNAs	Liu et al. [45], Nagano et al. [46] and Bologna and Voinnet [39]
	Formation of RNAi complex RISC	Bologna and Voinnet [39]
	Endogenous siRNA in plants: hp.-siRNAs, nat.-siRNAs	Borges and Martienssen [47], Chapman and Carrington [48] and Vasquez [49]
	Secondary siRNAs: tasiRNAs, phasiRNAs, easiRNAs	Borges and Martienssen [47] and Liu et al. [45]

2.1 *Types of Alternative Splicing (AS)*

Using different splice sites, the AS generates two or more mRNAs from the same pre-mRNA. Based on the type of AS, sequences of exons and introns are either included or excluded from the mRNA. The cassette exon is an exon that is included or excluded from the mRNA. Mutually exclusive exons refer to the splicing of the adjacent exon, causing only one of them to be included at a time in the mRNA. The alternative 5' splice site involves the use of the distal or proximal 5' splice site producing mRNAs of different size. The alternative 3' splice site involves the exploitation of the 3' proximal and distal splice sites, resulting in the production of mRNAs of different sizes. The final type of AS is the retention of an intron where the intron is retained or removed from the mRNA [27].

2.2 *Coupling of Transcription with AS*

More recently, it has become widely accepted that AS is a co-transcriptional event in which crosstalk is involved [33, 50]. The carboxy-terminal domain (CTD) involved in the coupling of transcription and processing steps is required for the recruitment of Ser/Arg-rich splicing factor 3 (SRSF3), which then inhibits the inclusion of alternative exons [51]. The mediator joins with the general transcription factors (GTFs) at promoters and specific TFs that are bound to gene enhancers, and recruits the negative splicing factor hnRNPL. This process causes hnRNPL to inhibit the inclusion of an alternative exon during splicing [52].

2.3 *AS in Plants*

AS is uncommon in unicellular eukaryotes and commonly found in multicellular eukaryotes and differs greatly among tissues and species [28, 53]. Only about 4% of the genes in budding yeast contain introns and AS is uncommon [54]. In comparison, the RNA structures containing exon-intron precincts, spliceosome components, and other splicing factors are commonly found in plants [27]. However, splicing in plants is unique due to their shorter introns compared to animals. Furthermore, intron retention is a common method of AS in plants and they contain more genes encoding Ser/Arg-rich (SR) proteins. The pre-mRNA of the spliceosomal proteins, particularly SR proteins, which have a key role in spliceosome assembly and splicing regulation, are extensively spliced. The availability of plant genome and transcript sequence data has allowed the global analysis of AS in many plant species. Genome-wide analysis of AS has been performed in model plants such as *Arabidopsis* [55], *Brachypodium* [56], and in crop plants such as rice [57] and soybean [58]. These studies have shown that plant genes have one or more alternative transcript isoforms

(~20% of the genes) [59]. Studies have shown that nearly 61% of multiexonic genes in *Arabidopsis* and nearly 33% of rice genes are alternatively spliced [55, 57]. The AS of genes has been studied to understand their role in plant growth development, environmental changes, and stress responses [60–62]. The studies mentioned above and others support the importance of intron retention in plants.

2.4 Database Resources of Plant Spliceosomal Proteins and AS

There are several resources available in relation to plant spliceosomal proteins and AS. Many of them are based on the model plant, *Arabidopsis*, such as the *Arabidopsis* slicing-related genes (ASRG) [63], and The *Arabidopsis* Information Resource (TAIR) [64]. Others include the AS in plants (ASIP), which is available for *Arabidopsis* and rice [65].

2.5 Role in Plant Development

Whole transcriptome profiling using RNA-seq was useful in enhancing the understanding of the gene expression of key genes and the coordinated expression of related genes during early somatic embryogenesis in maize [66]. AS is important in photosynthesis as it generates two protein products from the Rubisco activase gene, which is a nuclear-encoded chloroplast protein that mediates light activation of ribulose 1,5-biphosphate carboxylase/oxygenase (Rubisco) [67]. Based on the cDNAs and ESTs of *Arabidopsis* and rice analyzed using genome-wide computational analysis, AS has been shown to be common during flowering [65]. The alteration from the vegetative to the reproductive developmental stages is regulated by the alternative processing of the *FCA* pre-RNA [68]. Further, AS of the transcripts controls the spatial and temporal production of the FCA protein that regulates flowering.

2.6 Role in Biotic and Abiotic Stress Response

The pre-mRNAs of spliceosomal proteins are severely affected by biotic and abiotic stresses leading to AS [69, 70]. For example, *OSDREB2B* was shown to be regulated by stress-inducible AS [71]. Furthermore, the AS of the NRR transcript (related to root growth) identified in rice, produced two 5' co-terminal transcripts (NRRa and NRRb), and both products were shown to possess negative regulatory roles [72]. In another example, the TIR-NBS-LRR gene that is involved in tobacco mosaic virus

(TMV) resistance produced two transcripts, N_S and N_L , through AS. Expression of both transcripts were shown to be required for complete resistance to the virus [73]. Similarly, the combined presence of *RPS4* transcripts containing both full-length and truncated open reading frames was required to mediate disease resistance [74]. Abiotic stresses have been shown to affect the AS of the pre-mRNA in several SR genes [75]. The AS regulators such as the SR proteins, hnRNPs, and protein kinases have been suggested to play significant roles in stress responses [27]. They have been suggested to allow plants to react promptly in regulating splicing and gene expression.

2.7 *Alternative Polyadenylation*

The regulatory role of polyadenylation in eukaryotic gene expression involves alternative polyadenylation (APA) sites that produce different transcripts with altered coding capacity for proteins and/or RNA [34]. APAs have been reported in plants, in relation to flowering time control pathways [76], seed dormancy [77], and stress responses [78, 79]. It has been exhibited through global profiling methods that plants exploit APA for diversity generation in their transcriptomes. Through genome-wide analysis in *Arabidopsis*, the HLP1 protein was identified to regulate the pre-mRNA 3'-end processing and targets APA. It was enriched at transcripts involved in metabolism and flowering [76]. Another genome-wide study in *Arabidopsis* showed that the CPSF30 is associated with APA in response to oxidative stress [80]. The Plant APA is a recently developed database for the visualization and analysis of APA [81]. The role of AS in plant development and response to biotic and abiotic stresses is summarized in Table 2.

2.8 *Modifications in mRNA*

Recently, modifications of mRNA with N⁶-methyladenosine (m⁶A), 5-methylcytosine (m⁵C), and pseudouridine (ψ) have been revealed through technical advances. The m⁶A mRNA was the first internal mRNA modification to be identified. Due to its abundance it has been easily detected through bulk mRNA analysis, and NGS approaches have allowed the mapping of its locations [36]. Recently, transcriptome-wide m⁶A profiling of rice callus and leaf [35] and shoot [37] have been reported. There are, however, other RNA methylation events that have been found in other organisms and are thought to occur in plants. Subsequent studies will reveal their frequency of occurrence in plants and if they have any role in development or response to biotic/abiotic stresses.

Table 2 Role of alternate splicing, miRNA and siRNAs in plant development, and response to biotic and abiotic stresses

Response type	Function of mRNA modification, miRNA and siRNA in plants	Plant species	References
Plant development	Expression of genes during somatic embryogenesis	Maize	Salvo et al. [66]
	AS involved in photosynthesis	Rice	Zhang and Komatsu [67]
	AS involved in flowering	Arabidopsis and rice	Wang and Brendel [65]
	Alteration from vegetative to reproductive stages involving the <i>FCA</i> pre-RNA	Arabidopsis	Razem et al. [68]
	Pre-mRNA 3'- end processing and targeted APA in flowering and metabolism	Arabidopsis	Zhang et al. [76]
	APA in seed dormancy	Arabidopsis	Cyrek et al. [77]
Abiotic stress	<i>OSDREB2B</i> regulation by AS under drought and heat shock	Rice	Matsukura et al. [71]
	AS of <i>NRR</i> transcript under macro-nutrient deficiency	Rice	Zhang et al. [72]
	Association of <i>CPSF30</i> with APA under oxidative stress	Arabidopsis	Thomas et al. [80]
	miR319 expression increased salt and drought tolerance	Transgenic creeping bentgrass	Zhou et al. [82] and Zhou and Luo [83]
	miR319a/b, and miR319b.2 in copper, cadmium & sulphur deficiency conditions and salt stress	Arabidopsis	Barciszewska-Pacak et al. [84]
	miR169 repression under drought stress, phosphate deficiency, and nitrogen starvation	Arabidopsis	Hsieh et al. [85], Li et al. [86], Xu et al. [87] and Zhao et al. [88]
	miR169 repression under nitrogen-starvation	Maize	Xu et al. [87]
	Overexpression of miR169 under drought stress	tomato	Zhang et al. [89]
	Downregulation of mi169 and overexpression of StNF-YA genes enhanced drought tolerance	tomato	Yang et al. [90]
	Repression of <i>P5CDH</i> expression by nat-siRNA under salt stress	Arabidopsis	Borsani et al. [91]
siRNAs: <i>siRNA</i> 002061_0636_3054.1, 005047_0654_1904.1, 080621_1340_0098.1, 007927_0100_2975.1 were differentially expressed under cold heat, salt, and drought stress	Wheat	Yao et al. [92]	

(continued)

Table 2 (continued)

Response type	Function of mRNA modification, miRNA and siRNA in plants	Plant species	References
Biotic stress	AS involved in the TIR-NBS-LRR gene expression under TMV resistance	Tobacco	Dinesh-Kumar and Baker [73]
	AS in RPS-mediated disease resistance	Arabidopsis	Zhang and Gassmann [74]
	miR156, miR159, miR172, miR319, and miR393 responsive to <i>Cucumber mosaic virus</i>	Tomato	Feng et al. [93]
	Negative correlation between miR319 and its target TCP4 in response to RKN		Zhao et al. [94]
	miR319 responsive to <i>Verticillium longisporum</i>	Rapeseed	Shen et al. [58]
	miR393 responsive to <i>Pseudomonas syringae</i>	Arabidopsis	Navarro et al. [95]
	nat-siRNAATGB2 induced resistance against <i>Pst</i>	Arabidopsis	Katiyar-Agarwal et al. [96]

2.9 Stress Response Mechanism and the Cytoplasmic RNA-Containing Granules

It is important in transcriptomics to study the mRNAs that are translated, degraded, or stored temporarily during stress [97]. Based on the environmental or developmental conditions, the messenger ribonucleoprotein complexes (mRNPs) are formed through transcribed mRNA. While the polysome-associated mRNAs are translated, the non-translated mRNAs are localized on either the mRNA processing body (PB) or stress granules (SG), which are cytoplasmic mRNP granules. The PB (identified in yeast and mammals) contains RNA decay machinery for destroying unwanted mRNA in the 5'-3' direction. The SG store the non-translated mRNA that is stalled during initiation of translation and under stress conditions that cause the SG numbers to increase and accumulate. Several studies have suggested that SGs and PBs are an essential cytoplasmic structure that control gene expression during plant stress responses [98, 99].

3 microRNAs (miRNAs) and Small Interfering RNAs (siRNAs)

microRNAs (miRNAs, 19–25 nt) and small interfering RNAs (siRNAs, 21–22 nt) are small non-coding RNAs with important regulatory functions. Though miRNAs and siRNAs share a number of features in size, structure, and molecular function, they differ in biogenesis pathway and precursor structure [39, 49, 100, 101]. Both,

miRNAs and siRNAs are capable of producing a gene silencing effect at the post-transcriptional and transcriptional (epigenetic regulation) levels [38, 47, 102]. In contrast to miRNAs that are derived from either double-stranded or hairpin-like (60–70 nt) RNA precursors in almost all eukaryotes [38], siRNAs are generated from long double-stranded RNAs [103]. The miRNAs are endogenous, encoded by the host genome, while siRNAs can be exogenous or endogenous in origin. The former is originally derived from the transcription of viruses, transposons, repetitive DNA sequences, or transgene trigger [104, 105]. The miRNAs have numerous targets and regulate the expression of large numbers of target mRNAs. In contrast, the siRNAs are specific and mostly regulate the same genes they originate from [106]. Another major difference between miRNAs and siRNAs is that siRNAs base-pair to their target gene and exert targeted gene knockdown through the siRNA-induced mRNA cleavage, translational repression, and DNA methylation, whereas the former are partially complementary to target mRNAs and mediate post-transcriptional gene regulation through either mRNA cleavage or translational repression [47, 106].

3.1 Biogenesis of miRNAs in Plants

miRNAs are evolutionary highly conserved RNA molecules [39]. In the nucleus, miRNAs are transcribed by RNA polymerase II into primary miRNA (pri-miRNA), which are capped and polyadenylated. Subsequently, the pri-miRNAs are processed by the ribonuclease III enzyme, Dicer like 1 (DCL1), in the Dicer family, to a smaller stem-loop structure called precursor miRNAs (pre-miRNAs) [38, 39]. The pre-miRNAs are further processed again by DCL1 into the mature miRNA: miRNA* duplexes that carry 5' phosphates and 2-nt overhangs on their 3' end that are not fully complementary [40]. Next, they are exported from the nucleus to the cytoplasm by HASTY, the *Arabidopsis* homolog of exportin 5 in animals [41, 42]. In the cytoplasm, the mature miRNA strand, the so-called guide strand, is subsequently incorporated into the RNA-induced silencing complex (RISC, or miRISC for miRNA-containing RISC), where it is loaded onto a member of the Argonaute protein family and guides effector RISC to the target mRNA. The miRNA*, which is derived from the other strand known as the passenger strand, can be either degraded or functional [43, 44]. The key studies on miRNA and siRNA biogenesis are summarized in Table 1.

3.2 Biogenesis of siRNAs in Plants

In contrast to miRNAs, siRNAs are derived from double-stranded RNAs originating from protein-coding genes, non-coding transcripts, and transposable elements with perfect base-pairing complementarity to target mRNAs [47]. The DCL2, DCL3, and DCL4 sequentially cleave the long duplex structure into 22-nt, 24-nt, and 21-nt

siRNAs, respectively [39, 45, 46]. Short RNA duplexes are similar to the miRNA: miRNA* duplexes but are fully based-paired along the length. Once small RNA duplexes are generated, they are also loaded on an Argonaute protein. Next, the passenger strand is removed and the remainder forms the effector RNAi complex RISC (siRISC, which is loaded with siRNA [39]). Besides the RNA-dependent RNA polymerase 2 and 6 (RDR2, RDR6), SUPPRESSOR OF GENE SILENCING 3 (SGS3), and dsRNA-BINDING 4 (DRB4) are also implicated in siRNA biogenesis [107, 108].

Endogenous siRNAs in plants have been characterized based on their characteristics and biogenesis pathways into hairpin-derived siRNAs (hp-siRNAs, 21–24 nt), natural antisense siRNAs (nat-siRNAs, 21–24 nt), secondary siRNAs, and heterochromatic siRNAs (het-siRNAs, 24 nt) [47–49]. Secondary siRNAs could be subclassified into *trans*-acting siRNAs (tasiRNAs), phased siRNAs (phasiRNAs), and epigenetically activated siRNAs (easiRNAs) [45, 47].

3.3 *Functions of miRNAs and siRNAs in Plants*

A large number of miRNAs have been identified and characterized in plant genomes with diverse functions. Several miRNAs have been identified to have a crucial regulatory role in a wide range of biological processes in diverse plant species including but not limited to, cell signaling, differentiation, heterosis, DNA damage repair, hormone signaling, organ development, and response to biotic and abiotic stresses (reviewed in [47, 109, 110]).

The siRNA are widespread and numerous endogenous siRNAs have been discovered in plants by deep sequencing. The results of several studies revealed that siRNA is implicated in the morphological control of leaf [111]; developmental timing (temporal regulation) [112, 113]; hormone signaling [114]; fertility and reproductive function [115]; maintenance of genomic integrity, and developmental patterning [116].

Furthermore, both miRNAs and siRNAs, as part of multi-layered sophisticated defense mechanisms, play pivotal roles in regulating immune responses to environmental stresses [109, 116–118]. The role of miRNAs and siRNAs in plant development, and response to biotic and abiotic stresses is summarized in Table 2.

3.4 *Role of miRNAs in Plant Stress Responses*

Plants as sessile organisms are continuously and simultaneously challenged by multiple biotic and abiotic stressors. They have evolved sophisticated defense mechanisms and intricate regulatory networks to perceive their attackers. The miRNAs as critical regulators of gene expression, fine-tune defense responses by regulating the expression of their stress/defense-related target genes. Thousands of

miRNAs responsive to biotic and abiotic stresses and their targets have been identified in diverse plant species using deep sequencing technologies and degradome sequencing, respectively [119].

3.5 miRNAs in Biotic Stress

miRNAs orchestrate plant adaptive response to pathogens as the key players in hormone signaling pathways and plant immunity [89, 95, 117, 120]. Thus far, numerous biotic stress-responsive miRNAs have been identified in different plants [93, 94, 120–122]. Recently, several miRNAs such as miR156, miR159, miR172, miR319, and miR393 were found to be responsive to *Cucumber mosaic virus* in tomato [93]. In response to the fungal pathogen *Verticillium longisporum*, several miRNAs including the miR319 family have been identified in oilseed rape (*Brassica napus*) [121]. Flagellin-22 triggered miR393 expression and conferred resistance to *Pseudomonas syringae* in *Arabidopsis* [95]. In line with this, several miRNAs have been identified to be differentially expressed in *Arabidopsis* in response to a bacterial pathogen, *P. syringae* pv. tomato, using deep sequencing [122]. The results of these studies indicate that miRNAs target genes that are related to hormone signaling pathways and negatively regulate their target genes to enhance plant resistance to bacterial infection. The miR393 was reported to target *TIR1* (Transport Inhibitor Response 1) and its functional paralogs, *AFB2* and *AFB3* (Auxin signaling F-Box proteins 2 and 3). Whereas miR160 and miR167 target ARF8, ARF10, ARF16, and ARF17 to repress auxin signaling. Therefore, miRNAs confer a high degree of resistance to the bacterial pathogen *P. syringae* through miRNA-mediated suppression of auxin signaling [95, 122]. A recent study found that there is a negative correlation between miR319 and its target TEOSINTE BRANCHED1/CYCLOIDEA/PROLIFERATING CELL FACTOR 4 (*TCP4*) in response to root-knot nematode (RKN, *Meloidogyne incognita*) invasion in tomato (*Solanum lycopersicum* var Castlemart) [94]. The *TCP* genes encode plant-specific transcription factors that positively regulate jasmonic acid (JA) biosynthesis genes and JA levels in plants. The expression of miR319b was repressed, while the expression of its target, *TCP4*, was increased under JA treatment. On the other hand, the expression levels of all miR319-targeted *TCP* genes were significantly decreased in transgenic tomato plants overexpressing miR319 [94]. The results of this study showed that miR319 negatively regulates RKN resistance and JA-mediated miR319 confers systemic resistance to RKN infection. Additionally, crosstalk between miRNAs and hormone signaling pathways was revealed.

Altogether, miRNAs are responsive to a broad range of biotic stresses and confer resistance to plants against pathogens through complex mechanisms such as miRNA-mediated hormone signaling and/or hormone-mediated miRNA regulation.

3.6 miRNAs in Abiotic Stress

Several studies have shown that miRNA expression is regulated in response to a wide array of abiotic stresses such as drought, salinity, cold, heat, heavy metals, nutrients, oxidation, hypoxia, and UV-B in an miRNA-, stress-, tissue-, and genotype-dependent manner (reviewed in [109, 118, 119, 123]). The miR319 miRNA family is one of the most conserved and ancient miRNA families in plants [83]. miR319 was found to be induced in response to not only different biotic stresses, e.g., bacteria, fungi, viruses, and nematodes [93, 94, 121, 122], but also to multiple abiotic stress factors such as drought, salinity, cold, and aluminum [82, 83, 124–127]. Constitutive expression of miR319 significantly increased salt and drought tolerance in transgenic creeping bentgrass (*Agrostis stolonifera*) [82, 83]. Hence, miR319 can be a general multi-stress responsive miRNA. A recent study in *Arabidopsis* revealed that three miRNAs from the miR319 family, i.e., miR319a/b and miR319b.2, are associated with several abiotic stresses [84]. Interestingly, miR319a and miR319b exhibited the same patterns of expression in response to copper, cadmium, and sulfur deficiency conditions as well as salt stress. Moreover, the expression of miR319b.2 was augmented in response to copper, cadmium, and sulfur deficiency stresses, whilst it was down-regulated in response to drought, heat, and salinity. Similarly, the expression levels of miRNA319a/b were increased under metal stresses. On the other hand, miRNA319a/b was notably up-regulated under salinity stress [84]. These results suggest that miRNAs appear to have a complex regulatory role and orchestrate defense responses to a wide range of abiotic stresses through different regulatory networks.

In addition to the miR319 family, the miR169 family is another highly conserved family that plays a critical role in response to abiotic stresses in several plant species. The results of several studies indicated that miR169 plays an important role in response to several abiotic stresses including drought, salt, cold, abscisic acid, nitrogen starvation, and phosphate deficiency [85, 86, 88, 128–130].

The miR169 was repressed under drought and phosphate deficiency in *Arabidopsis* and nitrogen-starvation in *Arabidopsis* and maize [85–88]. In contrast, miR169 was up-regulated in response to cold stress in different plant species [130]. Overexpression of miR169 enhanced drought tolerance in *Solanum lycopersicum* [122]. The miR169 family members are up-regulated in *Arabidopsis*, maize, and soybean under cold, drought, and salinity stresses [129]. Their results showed that stress-induced miR169 promotes early flowering by repressing the *AtNF-YA* transcription factor [129]. Conversely, a recent study in *Solanum tuberosum* exhibited that downregulation of miR169 enhanced drought resistance through over-expression of *StNF-YA* genes [90]. Nuclear factor Y (NF-Y) transcription factors are the main targets of miR169. NF-Y encodes a CCAAT-binding transcription factor [86]. These findings revealed that there is a negative correlation between the expression of miR169 and its target NF-YA genes, and the miR169 regulates negatively and/or positively their target expression at the post-transcriptional level to enhance stress tolerance in different plant species

[90, 129]. Taken together, these results suggest that multi-stress responsive miR169 may orchestrate the expression of its target genes in a host- and stress-dependent manner. Different signaling pathways are mediated by miR169 and there is a complex crosstalk between the miR169 family members and their target transcription factors.

3.7 *Role of siRNAs in Plant Stress Responses*

Several studies have indicated that natural antisense transcripts (NATs) play a vital role in the regulation of defense signaling pathways and are involved in the response to different environmental stimuli by orchestrating corresponding NAT mRNAs [91, 96, 131].

3.8 *siRNAs in Biotic Stress*

nat-siRNAATGB2, the first endogenous siRNA, was specifically expressed in *Arabidopsis thaliana* leaves challenged with a virulent form of the bacterial pathogen *Pseudomonas syringae* pv. *tomato* (*Pst*) carrying effector *avrRpt2* [96]. The nat-siRNAATGB2 and its antisense target PPRL were transcribed in the opposite direction and a negative correlation was observed between the nat-siRNA and its antisense target expression in response to *P. syringae* infection. nat-siRNAATGB2 induced resistance by repressing the expression of PPRL as a negative regulator of the RPS2-mediated resistance against *Pst* that triggered hypersensitive response (HR) and cell death by recognition of *PstavrRpt2* effector [96]. A novel class of siRNAs known as long siRNAs (lsiRNAs, 30–40 nt) was discovered by Katiyar-Agarwal and colleagues in 2007. The results of this study revealed that lsiRNAs are stress-induced and expressed by a bacterial infection. AtlsiRNA-1 was remarkably and specifically over-expressed in response to *Pst* carrying effector *avrRpt2*. Overexpression of AtlsiRNA-1 repressed the expression of its target AtRAPmRNA and induced resistance by silencing AtRAP as a negative regulator of plant defense [132].

Using deep sequencing, 17,000 unique siRNAs corresponding to cis-NATs have been found in *Arabidopsis thaliana* in response to biotic stress in the form of bacterial infection and abiotic stresses in the form of cold, drought, and salt [72]. The results of these studies suggest that siRNAs are stress-induced and regulate defense response in plants through the reprogramming of gene expression.

3.9 siRNAs in Abiotic Stress

A large number of nat-siRNAs have been identified in rice (*Oryza sativa* cv. *japonica*) in response to cold, drought, and salt [133]. In *Arabidopsis thaliana*, the 21-nt nat-siRNAs repressed the expression of Δ^1 -pyrroline-5-carboxylate dehydrogenase (*P5CDH*), a stress-related gene, through mRNA cleavage under salt stress. Down-regulation of the *P5CDH* led to proline accumulation. Proline is as an osmoprotectant and ROS quencher that helps to tolerate salt stress, although under-expression of *P5CDH* instigated increased ROS production [91]. One study has indicated that four siRNAs were differentially expressed in response to cold, heat, salt, and drought in wheat (*Triticum aestivum*) [92]. The siRNA 002061_0636_3054.1 was significantly repressed by heat, salt, and drought stress; 005047_0654_1904.1 was strongly over-expressed in response to cold, whilst down-regulated in response to heat, salt, and drought stress; 080621_1340_0098.1 was faintly induced by cold and repressed by heat but not by either salt or drought stress; and 007927_0100_2975.1 was down-regulated by cold, salt, and drought stress [92]. Further, their results revealed that the four siRNAs were preferentially expressed in spikes and uniformly expressed in leaves and roots [92]. Therefore, the results of these studies exhibited that nat-siRNAs respond to biotic and abiotic stress conditions in a stress-specific and developmental stage-dependent manner.

4 Conclusions and Future Prospects

Alternative splicing, miRNAs, and siRNAs play critical regulatory roles in modulating gene expression during plant growth and development, in response to biotic and abiotic stresses, and plant adaptation to an ever-changing environment. Several small regulatory molecules have been identified to have versatile functions in food and feed crops. Manipulating expression levels of miRNAs and siRNAs in economically important crop plants can be an effective strategy to improve desirable traits, stress tolerance, and resiliency in response to environmental stress and pathogen attack in plants. Therefore, miRNAs and siRNAs can be used as new targets for developing trait-improved crop plants and improving plant tolerance to stresses. Two powerful genome editing tools, TALENs and CRISPR/Cas, can be used for targeted genome editing and knockdown/knockout of small RNAs.

In addition to the identification of small regulatory molecules and their transcriptional profiling, it is indispensable for scientific communities to understand the regulatory mechanisms of small RNAs that orchestrate cellular functions and adaptation to environmental stresses to minimize the unintended side effects in modified plants.

References

1. World Population Prospect: The 2015 Revision (2015) United Nations, Department of Economic and Social Affairs, Population Division. [Online] Available at: <http://www.un.org/en/development/desa/news/population/2015-report.html>. Accessed 7 Aug 2017
2. Valin H, Sands RD, van der Mensbrugge D, Nelson GC, Ahammad H, Blanc E, Bodirsky B et al (2014) The future of food demand: understanding differences in global economic models. *Agric Econ* 45:51–67
3. Gilbert N (2014) Cross-bred crops get fit faster. *Nature* 513:292
4. Kissoudis C, van de Wiel C, RGF V, van der Linden G (2014) Enhancing crop resilience to combined abiotic and biotic stress through the dissection of physiological and molecular crosstalk. *Front Plant Sci* 5:207
5. Gepts P (2002) A comparison between crop domestication, classical plant breeding, and genetic engineering. *Crop Sci* 42:1780–1790
6. Nejat N, Cahill DM, Vadamalai G, Ziemann M, Rookes J, Naderali N (2015) Transcriptomics-based analysis using RNA-seq of the coconut (*Cocos nucifera*) leaf in response to yellow decline phytoplasma infection. *Mol Gen Genomics* 290:1899–1910
7. Nejat N, Mantri N (2017) Plant immune system: crosstalk between responses to biotic and abiotic stresses the missing link in understanding plant defence. *Curr Issues Mol Biol* 23:1
8. Parker RM, Barnes NM (1999) mRNA: detection by in situ and northern hybridization. *Methods Mol Biol* 106:247–283
9. Weis JH, Tan SS, Martin BK, Wittwer CT (1992) Detection of rare mRNAs via quantitative RT-PCR. *Trends Genet* 8:263–264
10. Bustin SA (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 25:169–193
11. Deepak SA, Kottapalli KR, Rakwal R, Oros G, Rangappa KS, Iwahashi H et al (2007) Real-time PCR: revolutionizing detection and expression analysis of genes. *Curr Genomics* 8:234–251
12. Nejat N, Vadamalai G, Dickinson M (2012) Expression patterns of genes involved in the defence and stress response of *Spiroplasmacitri* infected Madagascar Periwinkle *Catharanthus roseus*. *Int J Mol Sci* 13:2301–2313
13. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470
14. Harmer SL, Hogenesch JB, Straume M, Chang H-S, Han B, Zhu T et al (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* 290:2110–2113
15. Mantri NL, Ford R, Coram TE, Pang ECK (2007) Transcriptional profiling of chickpea genes differentially regulated in response to high-salinity, cold and drought. *BMC Genomics* 8:303
16. Mantri NL, Ford R, Coram TE, Pang ECK (2010) Evidence of unique and shared responses to major biotic and abiotic stresses in chickpea. *Environ Exp Bot* 69:286–292
17. Zik M, Irish VF (2003) Global identification of target genes regulated by APETALA3 and PISTILLATA floral homeotic gene action. *Plant Cell* 15:207–222
18. Mardis ER (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9:387–402
19. Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
20. Mortazavi A, William BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628
21. ENCODE (2015) Research applications and users meeting. Bolger Center, Potomac
22. Rands CM, Meader S, Ponting CP, Lunter G (2014) 8.2% of the human genome is constrained: variation in rates of turnover across functional elements classes in the human lineage. *PLoS Genet* 10:e1004525
23. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74

24. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J (2014) Multiple evidence strands suggest that there may be as few as 19000 human protein-coding genes. *Hum Mol Genet* 23:5866–5878
25. Nejat N, Mantri N (2017) Emerging roles of long non-coding RNAs in plant response to biotic and abiotic stresses. *Crit Rev Biotechnol* 20:1–13. <https://doi.org/10.1080/07388551.2017.1312270>
26. Isshiki M, Tsumoto A, Shimamoto K (2006) The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. *Plant Cell* 18:146–158
27. Reddy AS (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol* 58:267–294
28. Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* 14:153–165
29. Wang J, Smith PJ, Krainer AR, Zhang MQ (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res* 33:5053–5062
30. Wang Z, Burge CB (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14:802–813
31. Le Hir H, Andersen GR (2008) Structural insights into the exon junction complex. *Curr Opin Struct Biol* 18:112–119
32. Schaal TD, Hertel KJ, Reed R, Maniatis T (2005) Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc Natl Acad Sci U S A* 102:5002–5007
33. Dujardin G, Lafaille C, Petrillo E, Buggiano V, Lig A, Fiszbein A, MAG H, Moreno NN, Muñoz MJ, Alló M, Schor IE (2013) Transcriptional elongation and alternative splicing. *Biochim Biophys Acta* 1829:134–140
34. Xing D, Li QQ (2011) Alternative polyadenylation and gene expression regulation in plants. *Wiley Interdiscip Rev RNA* 2:445–458
35. Li Y, Wang X, Li C, Hu S, Yu J, Song S (2014) Transcriptome-wide N6-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification. *RNA Biol* 11:1180–1188
36. Gilbert WV, Bell TA, Schaening C (2016) Messenger RNA modifications: form, distribution, and function. *Science* 352:1408–1412
37. Shen L, Liang Z, Gu X, Chen Y, ZWN T, Hou X, Cai WM, Dedon PC, Liu L, Yu H (2016) N6-Methyladenosine RNA modification regulates shoot stem cell fate in Arabidopsis. *Dev Cell* 38:186–200
38. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297
39. Bologna NG, Voinnet O (2014) The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. *Annu Rev Plant Biol* 65:473–503
40. Jones-Rhoades WM, Bartel DP, Bartel B (2006) MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol* 57:19–53
41. Bollman KM, Aukerman MJ, Park MY, Hunter C, Berardini TZ, Poethig RS (2003) HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development* 130:1493–1504
42. Park MY, Wu G, Gonzalez-Sulser A, Vaucheret H, Poethig RS (2005) Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A* 102:3691–3696
43. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
44. Meister G (2013) Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet* 14:447–459
45. Liu YX, Wang M, Wang XJ (2014) Endogenous small RNA clusters in plants. *Genomics Proteomics Bioinformatics* 12:64–71

46. Nagano H, Fukudome A, Hiraguri A, Moriyama H, Fukuhara T (2014) Distinct substrate specificities of Arabidopsis DCL3 and DCL4. *Nucleic Acids Res* 42:1845–1856
47. Borges F, Martienssen RA (2015) The expanding world of small RNAs in plants. *Nat Rev Mol Cell Biol* 16:727–741
48. Chapman EJ, Carrington JC (2007) Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet* 8:884–896
49. Vazquez F (2006) Arabidopsis endogenous small RNAs: highways and byways. *Trends Plant Sci* 11:460–468
50. Fujiwara N, Masuda S, Shiki T (2012) mRNA biogenesis in the nucleus and its export to the cytoplasm. INTECH Open Access Publisher, London
51. de la Mata M, Kornblihtt AR (2006) RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat Struct Mol Biol* 13:973–980
52. Huang Y, Li W, Yao X, Lin QJ, Yin JW, Liang Y, Heiner M, Tian B, Hui J, Wang G (2012) Mediator complex regulates alternative mRNA processing via the MED23 subunit. *Mol Cell* 45:459–469
53. Lareau LF, Green RE, Bhatnagar RS, Brenner SE (2004) The evolving roles of alternative splicing. *Curr Opin Struct Biol* 14:273–282
54. Barrass JD, Beggs JD (2003) Splicing goes global. *Trends Genet* 19:295–298
55. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res* 22:1184–1195
56. Sablok G, Gupta PK, Baek JM, Vazquez F, Min XJ (2011) Genome-wide survey of alternative splicing in the grass *Brachypodium distachyon*: a emerging model biosystem for plant functional genomics. *Biotechnol Lett* 33:629–636
57. Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res* 20:646–654
58. Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong LA, Peng DL, Tian Z (2014) Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* 26:996–1008
59. Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res* 18:1381–1392
60. Ali GS, Reddy ASN (2008) Regulation of alternative splicing of pre-mRNAs by stresses. *Curr Top Microbiol Immunol* 326:257–275
61. James AB, Syed NH, Bordage S, Marshall J, Nimmo GA, Jenkins GI, Herzyk P, Brown JW, Nimmo HG (2012) Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *Plant Cell* 24:961–981
62. Kumar S, Asif MH, Chakrabarty D, Tripathi RD, Trivedi PK (2011) Differential expression and alternative splicing of rice sulphate transporter family members regulate sulphur status during plant growth, development and stress conditions. *Funct Integr Genomics* 11:259–273
63. Wang BB, Brendel V (2004) The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome Biol* 5:1
64. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A (2007) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* 36:D1009–D1014
65. Wang BB, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci* 103:7175–7180
66. Salvo SA, Hirsch CN, Buell CR, Kaeppler SM, Kaeppler HF (2014) Whole transcriptome profiling of maize during early somatic embryogenesis reveals altered expression of stress factors and embryogenesis-related genes. *PLoS One* 9:e111407
67. Zhang Z, Komatsu S (2000) Molecular cloning and characterization of cDNAs encoding two isoforms of ribulose-1, 5-bisphosphate carboxylase/oxygenaseactivase in rice (*Oryza sativa* L.) *J Biochem* 128:383–389

68. Razem FA, El-Kereamy A, Abrams SR, Hill RD (2006) The RNA-binding protein FCA is an abscisic acid receptor. *Nature* 439:290–294
69. Kaashyap M, Ford R, Bohra A, Kuvalekar A, Mantri N (2017) Improving salt tolerance of chickpea using modern genomics tools and molecular breeding. *Curr Genomics* 18:557–567. <https://doi.org/10.2174/1389202918666170705155252>
70. Nakaminami K, Matsui A, Shinozaki K, Seki M (2012) RNA regulation in plant abiotic stress responses. *Biochim Biophys Acta* 1819:149–153
71. Matsukura S, Mizoi J, Yoshida T, Todaka D, Ito Y, Maruyama K, Shinozaki K, Yamaguchi-Shinozaki K (2010) Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes. *Mol Gen Genomics* 283:185–196
72. Zhang YM, Yan YS, Wang LN, Yang K, Xiao N, Liu YF, YP F, Sun ZX, Fang RX, Chen XY (2012) A novel rice gene, NRR responds to macronutrient deficiency and regulates root growth. *Mol Plant* 5:63–72
73. Dinesh-Kumar SP, Baker BJ (2000) Alternatively spliced N resistance gene transcripts: their possible role in tobacco mosaic virus resistance. *Proc Natl Acad Sci* 97:1908–1913
74. Zhang XC, Gassmann W (2003) RPS4-mediated disease resistance requires the combined presence of RPS4 transcripts with full-length and truncated open reading frames. *Plant Cell* 15:2333–2342
75. Egawa C, Kobayashi F, Ishibashi M, Nakamura T, Nakamura C, Takumi S (2006) Differential regulation of transcript accumulation and alternative splicing of a DREB2 homolog under abiotic stress conditions in common wheat. *Genes Genet Syst* 81:77–91
76. Zhang Y, Gu L, Hou Y, Wang L, Deng X, Hang R, Chen D, Zhang X, Zhang Y, Liu C, Cao X (2015) Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. *Cell Res* 25:864
77. Cyrek M, Fedak H, Ciesielski A, Guo Y, Sliwa A, Brzezniak L, Krzyczmonik K, Pietras Z, Kaczanowski S, Liu F, Swiezewski S (2016) Seed dormancy in Arabidopsis is controlled by alternative polyadenylation of DOG1. *Plant Physiol* 170:947–955
78. Motion GB, Amaro TM, Kulagina N, Huitema E (2015) Nuclear processes associated with plant immunity and pathogen susceptibility. *Brief Funct Genomics* 14:243–252
79. Tao P, Huang X, Li B, Wang W, Yue Z, Lei J, Zhong X (2014) Comparative analysis of alternative splicing, alternative polyadenylation and the expression of the two KIN genes from cytoplasmic male sterility cabbage (*Brassica oleracea* L. var. *capitata* L.) *Mol Gen Genomics* 289:361–372
80. Thomas PE, Wu X, Liu M, Gaffney B, Ji G, Li QQ, Hunt AG (2012) Genome-wide control of polyadenylation site choice by CPSF30 in Arabidopsis. *Plant Cell* 24:4376–4388
81. Wu X, Zhang Y, Li QQ (2016) Plant APA: a portal for visualization and analysis of alternative polyadenylation in plants. *Front Plant Sci* 7:889
82. Zhou M, Li D, Li Z, Hu Q, Yang C, Zhu L, Luo H (2013) Constitutive expression of a miR319 gene alters plant development and enhances salt and drought tolerance in transgenic creeping bentgrass. *Plant Physiol* 161:1375–1391
83. Zhou M, Luo H (2014) Role of microRNA319 in creeping bentgrass salinity and drought stress response. *Plant Signal Behav* 9:e28700
84. Barciszewska-Pacak M, Milanowska K, Knop K, Bielewicz D, Nuc P et al (2015) Arabidopsis microRNA expression regulation in a wide range of abiotic stress responses. *Front Plant Sci* 6:410
85. Hsieh LC, Lin SI, Shih AC, Chen JW, Lin WY, Tseng CY, Li WH, Chiou TJ (2009) Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. *Plant Physiol* 151:2120–2132
86. Li WX, Oono Y, Zhu J, HeX J, JM W, Iida K et al (2008) The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. *Plant Cell* 20:2238–2251

87. Xu Z, Zhong S, Li X, Li W, Rothstein SJ, Zhang S, Bi Y, Xie C (2011) Genome-wide identification of microRNAs in response to low nitrate availability in maize leaves and roots. *PLoS One* 6:e28009
88. Zhao M, Ding H, Zhu JK, Zhang F, Li WX (2011) Involvement of miR169 in the nitrogen-starvation responses in Arabidopsis. *New Phytol* 190:906–915
89. Zhang W, Gao S, Zhou X, Chellappan P, Chen Z, Zhou X, Zhang X et al (2011) Bacteria-responsive microRNAs regulate plant innate immunity by modulating plant hormone networks. *Plant Mol Biol* 75:93–105
90. Yang J, Zhang N, Zhou X, Si H, Wang D (2016) Identification of four novel stu-miR169s and their target genes in *Solanum tuberosum* and expression profiles response to drought stress. *Plant Syst Evol* 302:55–66
91. Borsani O, Zhu JH, Verslues PE, Sunkar R, Zhu JK (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell* 123:1279–1291
92. Yao Y, Ni Z, Peng H, Sun F, Xin M, Sunkar R et al (2010) Non-coding small RNAs responsive to abiotic stress in wheat (*Triticum aestivum* L.). *Funct Integr Genomics* 10:187–190
93. Feng JL, Liu SS, Wang MN, Lang QL, Jin CZ (2014) Identification of microRNAs and their targets in tomato infected with Cucumber mosaic virus based on deep sequencing. *Planta* 240:1335–1352
94. Zhao W, Li Z, Fan J, Hu C, Yang R, Qi X, Chen H et al (2015) Identification of jasmonic acid-associated microRNAs and characterization of the regulatory roles of the miR319/TCP4 module under root-knot nematode stress in tomato. *J Exp Bot* 66:4653–4667
95. Navarro L, Dunoyer P, Jay F, Arnold B, Dharmasiri N, Estelle M, Voinnet O, Jones JD (2006) A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science* 312:436–439
96. Katiyar-Agarwal S, Morgan R, Dahlbeck D, Borsani O, Villegas A, Zhu JK, Staskawicz BJ, Jin HL (2006) A pathogen-inducible endogenous siRNA in plant immunity. *Proc Natl Acad Sci USA* 103:18002–18007
97. Urano K, Kurihara Y, Seki M, Shinozaki K (2010) ‘Omics’ analyses of regulatory networks in plant abiotic stress responses. *Curr Opin Plant Biol* 13:132–138
98. Goeres DC, Van Norman JM, Zhang W, Fauver NA, Spencer ML, Sieburth LE (2007) Components of the Arabidopsis mRNA decapping complex are required for early seedling development. *Plant Cell* 19:1549–1564
99. Weber C, Nover L, Fauth M (2008) Plant stress granules and mRNA processing bodies are distinct from heat stress granules. *Plant J* 56:517–530
100. Mattick JS, Makunin IV (2006) Non-coding RNA. *Hum Mol Genet* 15:R17–R29
101. Zamore PD, Haley B (2005) Ribo-genome: the big world of smallRNAs. *Science* 309:1519–1524
102. Rogers K, Chen X (2013) Biogenesis, turnover, and mode of action of plant MicroRNAs. *Plant Cell* 25:2383–2399
103. Eamens A, Smith NA, Curtin SJ, Wang MB, Waterhouse PM (2009) The *Arabidopsis thaliana* double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *RNA* 15:2219–2235
104. Carthew R, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell* 136:642–655
105. Pontier D, Yahubyan G, Vega D, Bulski A, Saez-Vasquez J, Hakimi MA, Lerbs-Mache S, Colot V, Lagrange T (2005) Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev* 19:2030–2040
106. Mack GS (2007) MicroRNA gets down to business. *Nat Biotechnol* 25:631–638
107. Hiraguri A et al (2005) Specific interactions between Dicer-like proteins and HYL1/DRB-family dsRNA-binding proteins in Arabidopsis thaliana. *Plant Mol Biol* 57:173–188

108. Yoshikawa M, Peragine A, Park M-Y, Poethig RS (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. *Genes Dev* 19:2164–2175
109. Mantri N, Baskar N, Ford R, Pang ECK, Pardeshi V (2013) The role of micro-ribonucleic acids in legumes with a focus on abiotic stress response. *Plant Genome* 6:3
110. Wahid F, Shehzad A, Khan T, Kim YY (2010) MicroRNAs: synthesis, mechanism, function, and recent clinical trials. *Biochem Biophys Acta* 1803:1231–1243
111. Adenot X, Elmayan T, Lauressergues D, Boutet S, Bouche N, Gasciolli V, Vaucheret H (2006) DRB4-dependent TAS3 trans-acting siRNAs control leaf morphology through AGO7. *Curr Biol* 16:927–932
112. Poethig RS (2009) Small RNAs and developmental timing in plants. *Curr Opin Genet Dev* 19:374–378
113. Talmor-Neiman M, Stav R, Klipcan L, Buxdorf K, Baulcombe DC, Arazi T (2006) Identification of trans-acting siRNAs in moss and an RNA-dependent RNA polymerase required for their biogenesis. *Plant J* 48:511–521
114. Zubko E, Meyer P (2007) A natural antisense transcript of the *Petunia hybrida* Sho gene suggests a role for an antisense mechanism in cytokinin regulation. *Plant J* 52:1131–1139
115. Ron M, Saez MA, Williams LE, Fletcher JC, McCormick S (2010) Proper regulation of a sperm-specific cis-nat-siRNA is essential for double fertilization in Arabidopsis. *Genes Dev* 2010(24):1010–1021
116. Lelandais-Brière C, Sorin C, Declerck M, Benslimane A, Crespi M, Hartmann C (2010) Small RNA diversity in plants and its impact in development. *Curr Genomics* 11:14–23
117. Khraiweh B, Zhu JK, Zhu J (2012) Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochim Biophys Acta* 1819:137–148
118. Sunkar R, Li YF, Jagadeeswaran G (2012) Functions of microRNAs in plant stress responses. *Trends Plant Sci* 17:196–203
119. Zhang B (2015) MicroRNA: a new target for improving plant tolerance to abiotic stress. *J Exp Bot* 66:1749–1761
120. Baldrich P, Sun Segundo B (2016) MicroRNAs in rice innate immunity. *Rice (N Y)* 9:6. <https://doi.org/10.1186/s12284-016-0078-5>
121. Shen D, Suhrkamp I, Wang Y, Liu S, Menkhaus J, Verreet JA, Fan L, Cai D (2014) Identification and characterization of microRNAs in oilseed rape (*Brassica napus*) responsive to infection with the pathogenic fungus *Verticillium longisporum* using Brassica AA (*Brassica rapa*) and CC (*Brassica oleracea*) as reference genomes. *New Phytol* 204:577–594
122. Zhang X, Zou Z, Gong P, Zhang J, Ziaf K, Li H, Xiao F, Ye Z (2011) Over-expression of microRNA169 confers enhanced drought tolerance to tomato. *Biotechnol Lett* 33:403–409
123. Kumar R (2014) Role of microRNAs in biotic and abiotic stress responses in crop plants. *Appl Biochem Biotechnol* 174:93–115
124. Chen L, Wang T, Zhao M, Tian Q, Zhang WH (2012) Identification of aluminum-responsive microRNAs in *Medicago truncatula* by genome-wide highthroughput sequencing. *Planta* 235:375–386
125. Sunkar R, Zhu JK (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell* 16:2001–2019
126. Thiebaut F, Rojas CA, Almeida KL, Grativol C, Domiciano GC, Lamb CRC, Engler Jde A, Hemery AS, Ferreira PCG (2012) Regulation of miR319 during cold stress in sugarcane. *Plant Cell Environ* 35:502–512
127. Zhou L, Liu Y, Liu Z, Kong D, Duan M, Luo L (2010) Genome-wide identification and analysis of drought-responsive microRNAs in *Oryza sativa*. *J Exp Bot* 61:4157–4168
128. 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* 555:178–185
129. Xu MY, Zhang L, Li WW, Hu XL, Wang MB, Fan YL et al (2014) Stress-induced early flowering is mediated by miR169 in *Arabidopsis thaliana*. *J Exp Bot* 65:89–101

130. Zhang J, Xu Y, Huan Q, Chong K (2009) Deep sequencing of *Brachypodium* small RNAs at the global genome level identifies microRNAs involved in cold stress response. *BMC Genomics* 10:449
131. Jin H, Vacic V, Girke T, Lonardi S, Zhu JK (2008) Small RNAs and the regulation of cis-natural antisense transcripts in *Arabidopsis*. *BMC Mol Biol* 9:6
132. Katiyar-Agarwal S, Gao S, Vivian-Smith A, Jin H (2007) A novel class of bacteria-induced small RNAs in *Arabidopsis*. *Genes Dev* 21:3123–3134
133. Zhang X, Xia J, Lii YE, Barrera-Figueroa BE, Zhou X, Gao S et al (2012) Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. *Genome Biol* 13:R20

Metabolomics in Plant Stress Physiology



Arindam Ghatak, Palak Chaturvedi, and Wolfram Weckwerth

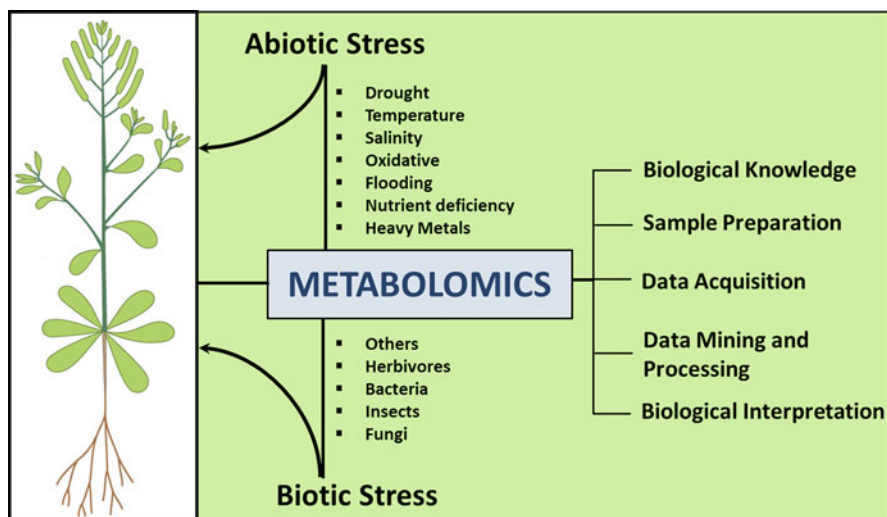
Abstract Metabolomics is an essential technology for functional genomics and systems biology. It plays a key role in functional annotation of genes and understanding towards cellular and molecular, biotic and abiotic stress responses. Different analytical techniques are used to extend the coverage of a full metabolome. The commonly used techniques are NMR, CE-MS, LC-MS, and GC-MS. The choice of a suitable technique depends on the speed, sensitivity, and accuracy. This chapter provides insight into plant metabolomic techniques, databases used in the analysis, data mining and processing, compound identification, and limitations in metabolomics. It also describes the workflow of measuring metabolites in plants. Metabolomic studies in plant responses to stress are a key research topic in many laboratories worldwide. We summarize different approaches and provide a generic overview of stress responsive metabolite markers and processes compiled from a broad range of different studies.

A. Ghatak and P. Chaturvedi
Department of Ecogenomics and Systems Biology, Faculty of Sciences, University of Vienna,
Vienna, Austria

W. Weckwerth (✉)
Department of Ecogenomics and Systems Biology, Faculty of Sciences, University of Vienna,
Vienna, Austria

Vienna Metabolomics Center (VIME), University of Vienna, Althanstrasse 14, 1090 Vienna,
Austria
e-mail: wolfram.weckwerth@univie.ac.at

Graphical Abstract



Keywords Abiotic stress, Analytical platforms, Biotic stress, Data mining, Functional genomics, Mass spectrometry, Metabolomics, Systems biology

Contents

1	Introduction	189
2	Analytical Platforms in Metabolomics	191
2.1	Nuclear Magnetic Resonance (NMR)	191
2.2	Mass Spectrometry (MS)	192
2.3	Capillary Electrophoresis (CE) MS	193
2.4	Data Mining and Data Processing	194
2.5	Compound Identification	195
2.6	Limitations of Metabolomics	196
3	Plant Metabolomics	197
4	Workflow for Plant Metabolomic Analysis	198
5	Metabolomic Studies in Plant Stress Responses	200
5.1	Drought Stress	200
5.2	Temperature Stress	211
5.3	Salt Stress	213
5.4	Oxidative Stress	215
5.5	Flooding	216
5.6	Nutrient Deficiency	216
5.7	Biotic Stress	219
5.8	Stress Combination	219
6	Metabolite Accumulation: Adjustment in Response to Stress Conditions	220
7	Conclusion and Outlook	222
	References	223

1 Introduction

Recent advances in technology have revolutionized the approach in which biological systems are visualized and questioned. Progressive developments in the field of genetics and automated nucleotide sequencing have supported the large-scale mapping and sequencing of many genomes, which include *Arabidopsis thaliana* [1], rice [2, 3], tomato [4], humans [5], etc. Technologies like expressed sequence tag (EST), mRNA profiling using microarrays [6], or serial analysis of gene expression (SAGE) [7] have allowed comprehensive analysis of the transcriptome. Advancements in mass spectrometry have enabled the analysis of cellular proteins and metabolites on a large scale, which was previously not possible [8–18]. The cumulative application of these technologies in various fields has led to advancement in the research of functional genomics and systems biology [19–23]. The foundations of both functional genomics and systems biology rely on comprehensive genome-scale molecular analysis [16, 18]. These approaches are commonly referred to as genomics, transcriptomics, proteomics, and metabolomics (Fig. 1).

Metabolomics is a complementary tool for functional genomics and systems biology, together with well established “omics” technologies for high-throughput

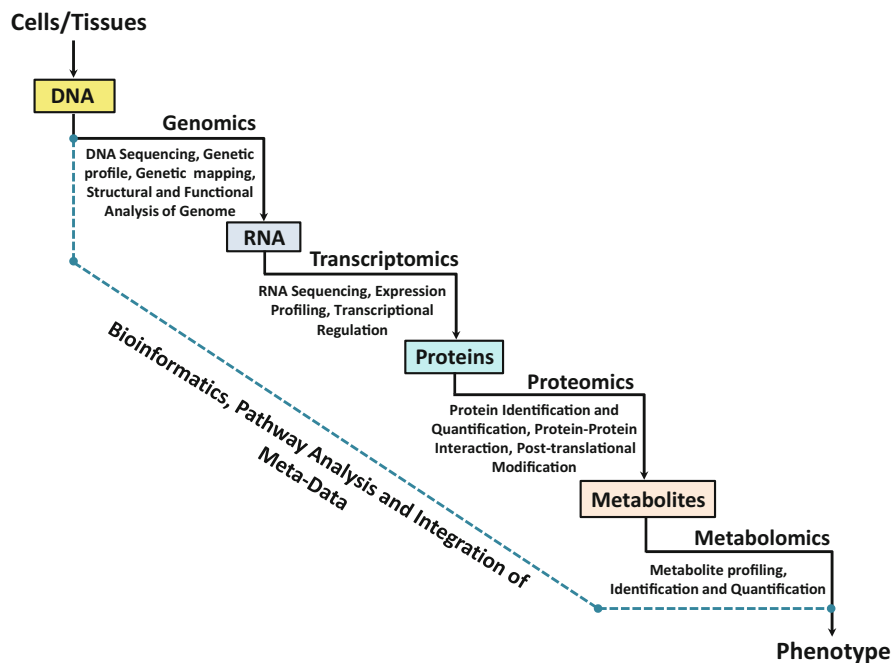


Fig. 1 Flow of biological information from genome to metabolome of a given cellular state in order to generate desired phenotype

data acquiring [16, 24–26]. The components of the metabolome can be viewed as the end-product of gene expression that defines the biochemical phenotype of a cell or a tissue. Quantitative and qualitative measurement of cellular metabolites thus provides a broad view of the biochemical status of an organism that can be used to monitor or assess gene function [16, 24]. The transcriptome represents mRNA changes in cellular machinery required for protein synthesis, but an increase in the levels of mRNA does not always correlate with protein levels [27]. Furthermore, translated proteins may or may not be enzymatically active. Thus, due to these reasons, changes in the transcriptome or proteome level do not necessarily correspond to the alteration in biochemical phenotypes. Moreover, transcriptome and proteome profiling includes the identification of mRNA and proteins through sequence similarity or database search (i.e., it depends on organism-specific genome information). Absence of these information/databases often limits the outcome of the analysis. Considering the above limitations, metabolite profiling provides essential functional information that has to be integrated with transcriptome and proteome analysis in order to increase the understanding of a given cellular state/biological sample [16–18, 28]. The qualitative and quantitative metabolomic profiling of a cell, tissue, or organism is very crucial because metabolites are structurally very small (less than 2,000 Da) and diverse molecules that are chemically transformed during cellular metabolism and hence pose a great challenge for analytical technologies [16, 29, 30]. Many stress responses lead to altered gene expression, particularly in plants, which results in qualitative changes in the metabolite pool, therefore identification of metabolites becomes even more critical [30].

One of the earliest metabolic profiling methods originated from Baylor College of Medicine in the early 1970s [31–33], which includes multicomponent analyses of steroids, acids, and neutral and acidic urinary drug metabolites using GC/MS. Thereafter, the concept of metabolite profiling was widely used for diagnostics and to assess health regimes [34, 35]. Gradually, there was an increase in research on automation [36] and expansion of GC-based methods to a wide range of chemical classes [37], followed by the use of high performance liquid chromatography (HPLC) and nuclear magnetic resonance (NMR) techniques for metabolite profiling [38, 39]. In the early 1990s, Sauter and colleagues used GC-MS metabolic profiling as the diagnostic technique in order to determine the mode of action of herbicides on barley plants [40]. Based on this and other studies, the concept of metabolic profiling and metabolomics in the context of functional genomics was introduced [24, 25, 41–43]. A first introduction of metabolomics as an integral technique for systems biology, linking metabolite profiling and metabolomics with genome-scale metabolic modelling, was described in 2003 [16]. There are many research institutes and commercial entities that are growing exponentially and working towards the development and improvisation of metabolomics science to broaden the range of its application.

2 Analytical Platforms in Metabolomics

A single analytical technique is not sufficient for detection and quantification of the metabolome and, therefore, multiple technologies are needed for a comprehensive view [16, 22, 29, 44]. Analytical technologies used in metabolomics include thin layer chromatography (TLC), HPLC with ultraviolet and photodiode array detection (LC/UV/PDA), gas chromatography–mass spectrometry (GC-MS), capillary electrophoresis–mass spectrometry (CE-MS), liquid chromatography–mass spectrometry (LC-MS), liquid chromatography–electrochemistry–mass spectrometry (LC-EC-MS), NMR, LC-NMR, direct infusion mass spectrometry (DIMS), and Fourier–transform infrared (FT-IR), etc. [16, 25, 43, 45, 46]. Of the above-mentioned techniques, NMR, GC-MS, LC-MS, and CE-MS are the most widely used technologies today [24, 45, 47–50]. Selection of the most suitable technology is based on speed, selectivity, sensitivity, and accuracy. NMR is rapid and selective, whereas mass spectrometry methods (GC-MS, LC-MS, and CE-MS) offer good selectivity and sensitivity but with longer analysis time [40, 51].

2.1 Nuclear Magnetic Resonance (NMR)

The use of NMR in metabolomics has opened the areas of biochemistry and phytochemical analysis [47, 52]. It is an unbiased, rapid, non-destructive technique that requires little sample preparation [48]. NMR analysis is not based on the analyte separation (as in the case of chromatographic analysis); rather it provides selectivity without separation and is also independent of the analyte polarity and does not require sample derivatization prior to the analysis. When the samples are placed in a strong magnetic field and irradiated with radio frequency, the absorption of energy promotes nuclei from a low-energy to a high-energy state. The subsequent emission of radiation generates resonance or signals that are recorded on the NMR spectrum as “chemical shifts,” representing frequencies from all NMR-visible nuclei in the sample, is relative to the reference proton present in a reference compound [53]. Hence, NMR analyses generally provide a global view of all the metabolites (primary and secondary) in a sample, provided they are in the detectable range [30, 49]. The disadvantage of low sensitivity and resolution has been addressed by using cryogenic probes, higher strength of superconducting magnets, miniaturized radio frequency coils, and multidimensional techniques (for example 2D-J-resolved and heteronuclear single quantum coherence) [48, 49]. 2D-NMR facilitates the identification of the compounds, which also includes the observation of minor compounds along with structural elucidation. Further, NMR has been applied in the areas of plant metabolism [54, 55], Duchenne muscular dystrophy [56], bio-availability, and metabolic responses of rats to epicatechin, hypertension, and acetaminophen toxicity [57, 58].

2.2 Mass Spectrometry (MS)

Flow-injection mass spectral analysis has been widely used in metabolic fingerprinting. Vaidyanathan et al. [59] demonstrated the use of flow injection ESI/MS for metabolic fingerprinting of cell-free extracts used for bacterial identification [59]. Similarly, multiple ionization techniques coupled to Fourier transform mass spectrometry (FT-MS) were used to identify metabolites associated with development and ripening of strawberry fruit [60]. FT-MS has a high resolution and high mass accuracy capacity, which allows separation and differentiation of very complex samples with the calculation of elemental composition, which facilitates structural differentiation and characterization. Unfortunately, this technology cannot differentiate chemical isomers that have an exact mass, e.g., hexoses [51].

Column chromatographic techniques (GC/LC) have a medium to high sensitivity that provides separation based on the physiochemical properties of an analyte [16, 50]. For complex samples, chromatographic separations include factors like column chemistry, an elution method for LC (gradient or isocratic), and a program-temperature method for GC. Multidimensional separation systems such as two-dimensional gas chromatography (GC \times GC) and two-dimensional liquid chromatography (LC \times LC) are used to enhance chromatographic separation of complex mixtures. Coupled to mass spectrometry (MS) for detection, these techniques are sensitive and capable of detecting low abundance metabolites [61–65]. GC-MS is a robust, technically reproducible, and sensitive approach. It is a well-suited platform for non-targeted metabolite profiling of volatile and thermally stable non-polar or derivatized polar metabolites [16, 26, 28, 66–69]. This technique is also used for targeted analysis of derivatized primary metabolites [70]. Electron impact (EI) is most commonly used in GC-MS, which results in fragmentation patterns that are highly reproducible. A mass analyzer like time-of-flight (TOF)-MS is widely used for detection because it has a faster “scan rate,” which improves deconvolution, high mass accuracy, and reduces the run time for complex mixtures [28, 66, 67, 69, 71, 72]. This technology is widely used for metabolite profiling, and thus contains several stable protocols for machine setup and maintenance, along with chromatogram evaluation and interpretation. Additionally, it has a short running time and a relatively low running cost. The use of the GC-MS platform is limited for thermally stable volatile compounds, thus making the analysis difficult for high molecular weight compounds (larger than 1 kDa) [16, 50, 73–75]. GC-MS facilitates the identification and quantification of hundreds of metabolites in plant samples, which include sugars, amino acids, organic acids, and polyamines, leading to the comprehensive coverage of primary metabolites in the central pathways. Quantitative metabolite profiling of potato tuber (*Solanum tuberosum*) using GC-MS leads to the identification of sugars, sugar alcohols, amino acids, and organic acids [69, 72, 76, 77]. The non-biased approach in the metabolite profiling of *Arabidopsis* leaf extract led to the identification of up to 652 metabolites and metabolite features [25, 28, 78]. Several studies were performed using other GC-MS methods, such as the metabolite profiling of tomato (*Lycopersicon esculentum*) and *Lotus japonicus*,

which led to the identification of 200 and 87 metabolites respectively [79, 80]. LC-MS does not require prior sample treatment and separates the components in a liquid phase [16, 50]. The choice of columns includes reversed phase ion exchange and hydrophobic interaction columns that separate metabolites based on the different chemical properties. In plant metabolomics, LC-MS is frequently used in the profiling of secondary metabolites [71, 81–85]. Hydrophilic interaction liquid chromatography coupled to mass spectrometry (HILIC-MS) is also widely used to analyze highly polar plant extracts [86, 87].

There are three main components in all types of MS instruments: (1) an ionization source such as electron impact (EI), electrospray (ESI), and atmospheric pressure chemical ionization (API); (2) a mass analyzer such as time of flight (TOF), quadrupole mass filters, and quadrupole ion trap; and (3) a detector such as an electron multiplier-based detector or micro-channel plate linked to a time-to-digital converter. The detected ions are recorded as pairs of m/z and abundance value, processed and displayed in a mass spectral format that allows us to identify and quantify a large variety of metabolites even with high molecular mass, high polarity, and low thermostability [45, 46].

2.3 Capillary Electrophoresis (CE) MS

Capillary electrophoresis (CE) uses charge-to-mass ratio to separate polar and charged compounds. CE is a powerful technique that can separate a diverse partially complementary range of chemical compounds compared to liquid chromatography [88–90]. In many CE-MS-based metabolomics studies, ESI is used for ionization in combination with TOF-MS, which provides high mass accuracy and high resolution. A small amount of a sample is required for the analysis (nanoliters of samples in the capillary). It can be used for volume-restricted sample analysis. One of the major drawbacks of this technique is less sensitivity, poor migration time, reproducibility, and lack of reference libraries. CE and LC can both separate a large variety of metabolites based on fundamentally different mechanisms, they can often be used together to provide a wider coverage of metabolites [91].

Watanabe et al. quantified carbohydrates, amino acids, and primary metabolites using CE-MS in response to elevated CO₂ [92]. Several studies have also been performed in rice using CE and CE-MS for the identification of primary metabolites. Maruyama and co-workers identified cold and dehydration-responsive metabolites, phytohormones, and gene transcription in rice. This analysis led to the identification of several genes that were up-regulated and involved in starch degradation, sucrose metabolism, and the glyoxylate cycle. It has also demonstrated the accumulation of glucose (Glc), fructose, and sucrose. Additionally, it was also observed that regulation of the glyoxylate cycle is correlated with glucose accumulation in rice, which was not observed in *Arabidopsis* [93].

2.4 Data Mining and Data Processing

All the “omics” analysis generates a large volume of data. In order to handle these large data sets, automated software is needed that can identify peaks from raw data, align the peaks among different samples, and replicate to identify and quantify each metabolite [81, 94–98]. Therefore, informatics and statistics are essential tools for processing metabolomics datasets. Data mining consists of data pre-processing, data pre-treatment, and statistical interpretation of the primary data [96]. The statistical interpretation is essential for central data analysis [68, 99, 100].

Metabolomics studies generate high-dimensional complex datasets that are difficult to analyze and interpret using univariate statistical analysis. Therefore, multivariate data analysis (MVDA) and mathematical modelling approaches are used to obtain meaningful information. These methods provide models that are well suited for a covariance pattern (both within and between variables) analyses [16, 17, 24, 28, 68, 98, 101–103]. Most commonly used methods for MVDA are principal component analysis (PCA), ANOVA, partial least square (PLS), and SIMCA (Soft Independent Modelling of Class Analogy). Web-based applications like MetaGeneAlyse implements standard normalization/clustering methods like k-means and independent component analysis (ICA) [104]. This software also provides other statistical analysis like the t-test, PLSDA (partial least square discriminant analysis), pathway enrichment analysis, etc. Other web-based applications are MetaboAnalyst [105], MetaMapp [106], metaP-Server [107], MeltDB [108], MetiTree [109], etc. These applications cover multiple steps from data pre-processing to biological interpretation. Many different multivariate statistical tools, metabolic modelling, and structural elucidation of unknown metabolites were integrated into a toolbox for metabolomics called COVAIN [96, 97]. The name stems from “covariance inverse,” implying that a covariance pattern is indeed used for functional interpretation of metabolite dynamics. This concept was recently developed and provides a fundamental novel approach to link causal relationships of biochemical networks with metabolite dynamics measured with metabolomics technology such as GC-MS, LC-MS, or any other technology [16, 18, 50, 68, 96, 97]. This approach goes beyond the classical MVDA analysis or correlation network analysis because it is able to identify causal biochemical perturbation points. It was recently applied to the analysis of the interface of primary and secondary metabolism in plants in *Arabidopsis* and *Theobroma* [81, 84, 85] and for the analysis of energy starvation and subsequent biochemical regulation [110, 111].

The high throughput metabolic data can be analyzed in supervised and unsupervised strategies [68]. The unsupervised method focuses on the intrinsic structure, relation, and interconnection of the data, which are sometimes referred to as descriptive and explorative models. Supervised methods seek to transform multivariate data from metabolite profiling into representations of biological interest under “supervision” and are often referred to as predictive models [29].

Metabolite data contain information such as metabolite name, change in levels, and their relationships, which are very useful for the interpretation of the biological significance. Most commonly, the identified metabolite is described in pathways or

Table 1 List of metabolomics databases involving tools for analysis, data processing, statistical analysis, biomathematical modeling and functional interpretation

Tools for analysis	Website URL
COVAIN	http://www.univie.ac.at/mosys/software.html
MetaboAnalyst	http://www.metaboanalyst.ca/
MetaMapp	http://metamapp.fiehnlab.ucdavis.edu/
MetiTree	http://www.metitree.nl/
MetaGeneAlyse	http://metagenealyse.mpimp-golm.mpg.de/
metaP-server	http://metabolomics.helmholtz-muenchen.de/metap2/
MeltDB 2.0	https://meltdb.cebitec.uni-bielefeld.de/cgi-bin
MetMask	http://metmask.sourceforge.net/
MetNetDB	http://www.metnetdb.org/MetNet_overview.htm
SMPDB	http://smpdb.ca/
GMD@CSB.DB: The Golm metabolome database	http://gmd.mpimp-golm.mpg.de/Default.aspx
McGill metabolome database	http://metabolomics.mcgill.ca/
SoyMetDB	http://soymetdb.org/
MoTo DB	http://www.transplantdb.eu/node/1843
Pathway-related databases	Website URL
MapMan	http://mapman.gabipd.org/
MetaCyc	http://metacyc.org/
MetaCrop	http://metacrop.ipk-gatersleben.de
AraCyc	https://www.arabidopsis.org/biocyc/
BioCyc	http://biocyc.org/
AraPath	http://bioinformatics.sdstate.edu/arapath/
KaPPA-view	http://kpv.kazusa.or.jp/
KEGG	http://www.genome.jp/kegg/
VANTED	https://immersive-analytics.infotech.monash.edu/vanted/
Pathvisio	http://www.pathvisio.org/
PlantCyC	http://www.plantcyc.org/databases

networks to understand its biological context. Some of the well curated databases for metabolic pathways in plants are KEGG (which is based on resources like GenBank/EMBL/DDJB) [112], ArcCyc [113], MetaCrop [114], UniPathway [115], SMPDB [116], and MapMan [117] etc. Table 1 shows the details of other metabolomic databases that are widely used for the analysis.

2.5 Compound Identification

Compound identification is the conclusive step in metabolite analysis; it is one of the critical steps because the biochemical interpretation of metabolomic data is based on

the availability of a well-structured database for identification of metabolites [118, 119]. Putative compound identification is based on molecular properties like the mass spectral pattern and accurate mass to define molecular and/or empirical formulae from which the metabolite can be derived or identified by the comparative search [82]. Definitive compound identification is based on retention time (Rt), retention index (RI), mass spectral fragmentation, and NMR spectral shift. Confirmation of the identified compound is done by a comparative library search, authentic chemical standards, and by using *in vivo* labelling methods. Sometimes analytically detected entities with biological significance are reported as “unknowns” with no structural identification [46, 118]. Recently, we have introduced a novel algorithm for structural elucidation of unknown compounds and even full pathways from untargeted metabolomics data [82]. This algorithm is especially suited for stress-related secondary metabolites such as flavonoids as an antioxidative response to cold and light stress [82]. In this study we also demonstrated how cold and light stress change the oxygen-to carbon-ratio in secondary metabolites systematically using so-called van Krevelen plots [82]. For compound identification at different levels of accuracy, minimum reporting standards need to be described. The metabolomics community has developed a nomenclature for publication of metabolomics data [120].

2.6 *Limitations of Metabolomics*

Metabolomic platforms lack the ability to comprehensively profile all the metabolites of a given cell/tissue [98]. This limitation is directly linked to the chemical complexity of the metabolites, the biological variance that is inherent in living organisms, and the dynamic range of the instruments. The genome and transcriptome consist of linear polymers of nucleotides with high chemical similarity; this structure facilitates high-throughput analytical approaches. The proteome is substantially more complex, but it is still based on a limited set of amino acids. The chemistry of these biopolymers is well defined and analytical technologies like 2DE gel electrophoresis and shotgun proteomics can readily identify and differentiate a large number of proteins in a single analysis and even post-translational modifications such as phosphorylation or methylation [14, 121, 122]. In the case of the metabolome, the chemical complexities are significantly greater and range from an ionic inorganic moiety to hydrophilic carbohydrates, hydrophobic lipids, and complex natural products. Hence, the chemical diversity and complexity make metabolome profiling extremely difficult. This obstacle can be circumvented by using selective extraction protocols and combinations of technologies for the analysis to obtain a more comprehensive coverage of the metabolome [16, 28, 71, 81].

Analytical variation can be defined as the coefficient of variation or relative standard deviation that is directly related to the experimental approach; this variance differs depending on the technology platform being employed. Biological variance arises from the quantitative variation in metabolite levels between plants of the same

species that are grown under identical conditions [28, 123]. Biological variance is the major limitation of “resolution” in metabolomics. Pooling of the samples tends to avoid or reduce biological variance. This strategy helps to minimize random variation by using statistical knowledge, but also leads to dilution of the samples, which results in the dilution of sites or tissues that are important for specific regulations (up/down) of the metabolite. Therefore, more targeted analysis can help to minimize the variation. In the case of plants, parameters like the synthesis of natural products, growth stage, environmental cues, etc. make the sampling critical and strategies are required to minimize the variations [44, 51].

The major analytical challenge encountered in metabolomics is dynamic range. Dynamic range can be defined as the concentration boundaries for analytical determination. The dynamic range can be critically limited by a sample matrix or by the presence of interfering and competing compounds. Most of the mass spectrometers have a dynamic range of 10^4 – 10^6 for individual components; however, this range is reduced significantly due to the presence of other chemical components (i.e., the presence of excessive metabolites that can cause significant interference that limits the range for the identification of other metabolites) [44, 51]. For example, high levels of sugars (primary metabolites) often interfere with the identification of secondary metabolites such as flavonoids. However, many of the highly expressed metabolites are often unique and provide a basis for the differentiation of the cellular states, organs, tissues, and organisms. These exclusive compounds are often referred as “biomarkers.” Selective profiling of these biomarkers is very useful for high-throughput diagnosis of a specific disorder, for example diabetes (i.e., glucose monitoring) or cancer. This detection should not be regarded or classified as metabolomics due to the highly targeted nature of profiling [24]. Another problem is salts; low levels of these ionic species reduce the ionization efficiency in ESI/MS and significantly interfere with the profiling of all other species [124]. Therefore, different analytical approaches have been developed to improve the dynamic range and to increase the level of identifications [51].

3 Plant Metabolomics

Metabolomic platforms can be used for unbiased identification of the metabolite levels in different genotypes that may or may not produce visible phenotypes [16, 72, 125]. The total number of metabolites found in plants are currently estimated to be ~200,000, with ~7,000–15,000 found in any individual species [73], of which 3,000–5,000 were exclusively determined in leaves [48]. So far metabolite profiling has been performed on a wide range of plant species, which includes *Arabidopsis*, tomato, potato, rice, wheat, strawberry, *Medicago*, cucumber, lettuce, tobacco, poplar, and Eucalyptus.

Selective metabolite profiling has been used in many studies to provide biological information beyond simple identification of the plant constituents. These include: (1) fingerprinting of species, genotypes, and ecotypes for taxonomic or biochemical

information [29, 44, 126]; (2) monitoring the behavior of a specific class of metabolites in response to exogenous chemicals or physical stimuli [127, 128]; (3) understanding the developmental process and symbiotic associations [129]; and (4) comparison of the metabolite content of mutant or transgenic plants with the wild type [44, 51]. In each of these studies, metabolite profiling can be coupled with other “omics” technologies to provide an integrated picture of all aspects of information from genome to metabolome and the resulting phenotype [18]. For example, metabolite profiling can be combined with marker-assisted selection providing integral information and understanding about the chemical composition of crop species [130]. Apart from the wide application of metabolomics in plant research, there has been considerable progress in the metabolomic analysis of single cells from different plant species, which includes tissues like pollen, trichome, root hairs, guard cells, etc.; these studies have been well reviewed [131].

The phenotype of a plant depends on the synthesis and accumulation of a series of metabolites in specific organs, at specific developmental stages [132]. It also depends on the environmental signals. Therefore, there are various kinds of metabolites in plants that have organ-/tissue-specific characteristics. For example, sphingolipids, a class of lipids that are critical in the development of the male gametophyte, are significantly different in the pollen and leaf tissues of *Arabidopsis* [133]. Anthocyanins accumulate in hypocotyls of young tomato seedlings, whereas several flavonols and phenolic compounds have been identified in cotyledons and some alkaloidal compounds are found in the root [132]. There are a lot of variations in biochemical pathways at cellular and sub-cellular levels in plants; therefore, the use of metabolomic platforms with different methods has significantly increased.

4 Workflow for Plant Metabolomic Analysis

Plant metabolomes are very complex and diverse in their chemical structures. Comprehensive identification and a broad range of metabolic pictures can be achieved by the combination of two or more metabolomic strategies and analytical systems, including variation in the extraction protocols [16, 28, 49, 71, 81, 134–136].

Metabolomic analyses consist of three main experimental strategies: (1) sample preparation; (2) acquisition of the data using analytical methods; and (3) compound identification and data mining. These steps are crucial and inter-related, as is illustrated in Fig. 2, with each step consisting of a series of sub-steps with various experimental phases to form a meaningful biochemical interpretation [118, 136].

Sample preparation is one of the critical steps as it contributes to the identification of the wide array of metabolites. This step consists of selection and harvesting of samples, drying or quenching procedure, and extraction of metabolites for analysis (derivatization). The selection of plant material depends on the researcher and the experimental design. Throughout this step, care must be taken to avoid the introduction of unwanted variability, which could significantly affect the outcome of the analysis. Sample degradation (oxidative or enzymological) and contamination are the major factors. Various enzyme quenching methods like drying, use of enzyme

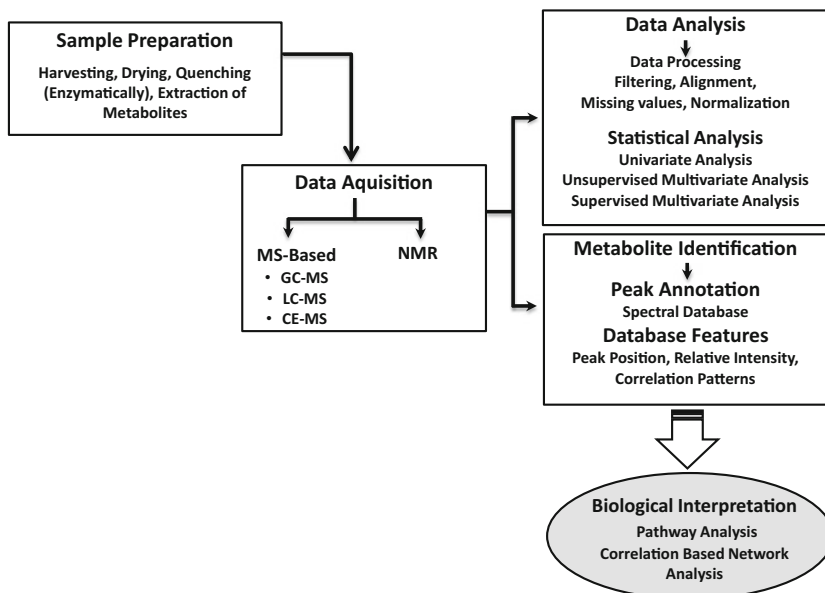


Fig. 2 Flowchart detailing the steps involved in plant metabolomics. There are three main steps in the metabolomic analysis, which include sample preparation, data acquisition, and data analysis, which lead to the biochemical interpretation of the identified metabolite

inhibitors, use of acids, or high concentrations of organic solvents can also affect the analyses/identification [24, 136].

Plant metabolites are structurally diverse with high complexities like different size, solubility, volatility, polarity, quantity, and stability [51]. There are several metabolite extraction protocols; the choice of method depends on a variety of factors such as physiochemical properties of the targeted metabolites, biochemical composition, and solvent used. Some of the common extraction protocols include solvent extraction, supercritical fluid extraction, solid phase extraction, and sonication [30, 52, 136]. However, no comprehensive extraction technique exists that can lead to the identification of all classes of metabolites with high reproducibility and robustness.

Sample analysis requires an advanced analytical platform (separately/combination) to measure the ultra-complex metabolite samples [51]. The ranges of analytical platforms are well discussed above, and each platform has its own limitations, either in sensitivity or selectivity. The choice of the platform depends on the study undertaken, consideration of the class of compounds, and their chemical and physical properties along with their concentration levels [45, 137].

Recently we have developed an integrative protocol that combines comprehensive metabolite extraction and analysis with proteomics and RNA analysis from one sample [28]. This protocol is adapted to allow the simultaneous analysis of all molecular levels and investigate their inter-relation and covariance structure [28, 66, 67, 138, 139]. This covariance structure of molecular dynamics of a cellular

system is a result of biochemical regulation [139]. Therefore, it is possible to read biochemical regulation from the molecular association networks or in other words the corresponding covariance data [16–18, 66, 68, 72, 84, 110, 139].

5 Metabolomic Studies in Plant Stress Responses

Metabolomic studies have become increasingly common in plant physiology and biochemistry. In the following section, we review applications of metabolomics to study plant responses to environmental stress (like abiotic and biotic factors) (Table 2). Abiotic factors include drought, temperature, salt, oxidative stress, flooding, nutrient deficiency, heavy metals, and effects of combinations of stress. We describe the nature and symptoms of various stresses and introduce several metabolomic studies with major metabolite changes (Table 2).

5.1 Drought Stress

Drought is one of the major threats to crop production worldwide and the situation is projected to get worse in the near future [193, 194]. The physiological response of plants under drought stress includes reduction of leaf area, leaf abscission, and increase in root growth to enhance the uptake of nutrients. Additionally, closure of leaf stomata takes place, which reduces water loss through transpiration. These physiological changes improve the water use efficiency (WUE) of the plant in the short term, but have a negative effect on photosynthesis, for example, closure of stomata lowers the intercellular CO₂ concentration, which adversely affects photosynthesis [193, 194]. Fine metabolomic adjustment is another strategy of plants to cope with drought stress. These metabolomic adjustments include net accumulation of osmolytes in the cell to retain or promote the uptake of water into the cell via osmosis in order to maintain turgor pressure [193].

Compatible solutes are highly soluble and small molecular-weight osmolytes that do not inhibit cellular metabolism even at high concentrations [195, 196]. Common osmolytes include soluble sugars (e.g., glucose, sucrose, and trehalose), the RFOs (raffinose, stachyose, and verbascose), polyols (e.g., mannitol, sorbitol), amino acids (e.g., proline), quaternary ammonium compounds (e.g., glycine betaine), and other polyamines (e.g., putrescine, spermidine, and spermine). Accumulation of these osmolytes in plant cells is important for sustaining cell turgor by osmotic adjustment, and stabilization of enzymes that reduce the levels of reactive oxygen species (ROS) in order to maintain the cellular redox balance.

The targeted metabolite approach using LC-MS was established by Antonio and co-workers in 2008 to investigate the effect of drought in *Lupinus albus* stem tissues [140]. In this analysis, 12 water-soluble organic osmolytes – like mono- and disaccharides, raffinose, stachyose, and verbascose – and sugar alcohols were identified by chromatographic analysis using PGC stationary phase. Although the

Table 2 List of significantly changed metabolites under stress conditions

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
Metabolites under drought stress conditions				
<i>Lupinus albus</i>	Stem	PGC-LC-ESI-MS	Raffinose, Stachyose, Verbascose, Sucrose, Glucose	[140]
<i>Haberlea rhodopensis</i>	Leaf	GC-MS, LC-MS	Maltose, Sucrose, Threonate, Fructose, Raffinose, Trehalose, Serine, Glycine, Myconoside, valine, Isoleucine, Phenylalanine, Tyrosine, Threonine, Asparagine, Proline, Glutamate, Malate, Phosphate, Fumarate, Spermidine, beta-Alanine	[141]
<i>Physcomitrella patens</i>	Multiple tissue	GC–quadropole MS	Altrose, Fructose, Glycopyranoside, Isomaltose, Maltitol, Ascorbic acid, Proline, Allonic acid, Galactonic acid, Myo-Inositol, Alanine, Glycine, Norvaline, Threonine, Aspartic acid, Glutamine, Serine, Threonine, Ascorbic acid, Butyric acid, Oxalic acid, Malic acid, Allantoin, Phosphoric acid	[142]
<i>Triticum aestivum</i>	Leaf	GC-MS	Glutamine, Quebrachitol, Proline, Methionine, Pipecolate, Lysine, 2-Amino-adipate, Asparagine, beta-Alanine, Homoserine, 2-Oxo-butanoate, Isoleucine, Valine, Leucine, Tyrosine, Oleic acid, Linoleic acid, Octadecanol, Nonanoate, Tryptophan, Phenylalanine, Rhamnase, Fucose, Mannose, Galactose, Sucrose, Gentibiose, Mannitol, Glycerol-3-P, Digalctosyl-glycerol, Xylose, Ribonate, Arabinose, Ribose, Galactinol, Inositol, Erythronate, Isocitrate, Succinate, Fumarate	[143]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Oryza sativa</i>	Leaf	GC/EI-TOF-MS	Putrescine, Spermidine, Spermine, Arginine, Ornithine, GABA, beta-Alanine, Proline, Glutamate	[144]
<i>Oryza sativa</i>	Leaf	HPLC	Putrescine, Spermidine, Spermine, Arginine, Ornithine	[145]
<i>Arabidopsis thaliana</i>	Aerial parts	GC/TOF-MS, CE-MS	Agmatine, Proline, Lysine, Methionine, Saccharopine, beta-Alanine, Phenylalanine, Galactinol, Raffinose, Citrate, Malate, Succinate, Alanine, GABA, Adenosine, Ascorbate, Glycine, Sarcosine, Threonine, Uracil, Xylose, Ascorbate, Pyruvate, Aspartate, Dehydro-L-ascorbate, Glucose, maltose, myo-Inositol, Indole-3-acetone, Succinimide	[146]
<i>Arabidopsis thaliana</i>	Leaf	GC/TOF-MS	Arginine, Ornithine, Proline, Putrescine, Erythritol, Fumarate, Fructose, Glucose, Aspartate, Glycine, myo-Inositol, Isoleucine, Trehalose, Valine, Glycerate, Glycerol-3-P, malate, Threonate, Threonic acid, Alanine, Arginine, Fructose-6-P, Fumarate, Glutamate, Glutamine, 2-oxoglutarate, Homoserine, Maltose, Methionine, Ornithine, Phenylalanine, Pyroglutamate, Serine, Threonine	[147]
<i>Solanum lycopersicum</i>	Fruit, pericarp tissue	GC-MS	Aconitate, Citrate, Isocitrate, Fumarate, Malate, Succinate, Sucrose, Inositol, Tocopherol, Phenylalanine, Methionine, Dehydroascorbate, Alanine, Aspartate, beta-Alanine, GABA, Glutamate, Glycine, Homoserine, Isoleucine, Proline, Serine, Valine, Cysteine	[148]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Zea mays</i>	Multiple tissues	GC-TOF-MS	Xylose, Vanillic acid, Methionine, Putrescine, Proline, Histidine, beta-Alanine, Phosphoric acid, Pyruvic acid, Succinic acid, Isoleucine, Phenylalanine, Adenine	[149]
<i>Zea mays</i>	Leaf	UPLC-GC/MS	Aconitate, Shikimate, Serine, Glycine, Proline, Malate, Fumarate, 2-oxoglutarate, Nitrate, Alanine, Aspartate, Sucrose, myo-Inositol, Raffinose, Shikimate, Glutamate, Threonine, GABA, Methionine, Tyrosine, Leucine, Phenylalanine	[150]
<i>Glycine max</i>	Leaf and nodule	NMR	Myo-Inositol, Pinitol, Sucrose, Alanine, Asparagine, Aspartate, Choline, Proline, Glutamine, GABA, Succinic acid, Fumaric acid, Malic acid, Citric acid, 2-Oxoglutarate	[151]
Metabolites under temperature stress				
<i>Arabidopsis thaliana</i>	Complete plant	GC-TOF-MS	Glutamine, Proline, Fructose, Galactinol, Glucose, Raffinose, Aconitate, alpha-ketoglutarate, Ascorbate, Gluconate, Gluconatelactone, Isosuccinate, Malate, Inositol, Melibiose, Threitol, Trehalose, Asparagine, Citrulline, Cycloserine, Glycine, Homoserine, Ornithine, Putrescine, Tryptophan, Proline	[152]
<i>Arabidopsis thaliana</i>	Complete rosettes	GC-TOF-MS	Lysine, Fructose, Sorbose, Glucose, Ornithine, Raffinose, Serine, beta-Alanine, Homoserine, Methionine, Arginine, Isoleucine, Valine, Glycine, GABA, Alanine, Phenylalanine, Pyruvic acid, Glutamine, Maltose, Trehalose, Proline, Asparagine	[153]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Arabidopsis thaliana</i>	Aerial tissues	GC-MS	<p>Heat Stress: Uracil, D-(2)-Quinic acid, Citramalic acid</p> <p>Cold Stress: Allantoin, Aconitic acid, Ferulic acid, Sinapic acid, Fru-6-P, Glyceric acid-3-P, Ascorbic acid, Arginine, Cysteine, Proline, Tryptophan, Maleic acid, Octadecanoic acid, Sorbitol, Isocitric acid, Lactic acid</p> <p>Common response: 2-Ketoglutaric acid, beta-Alanine, Citric acid, Erythritol, Fructose, Fumaric acid, GABA, Galactinol, Glycerol, Glycine, Alanine, Isoleucine, Valine, Putrescine, Trehalose, Xylose, Dehydroascorbic acid dimer, Glyceric acid, Asparagine, Succinic acid, Myoinositol-P, Ornithine</p>	[154]
<i>Arabidopsis thaliana</i>	Aerial tissues	GC-MS	<p>Proline, Glutamate, Glutamine, Arginine, Ornithine, Putrescine, GABA, Lysine, Methionine, Isoleucine, Threonine, Aspartate, Asparagine, Valine, Leucine, Alanine, Shikimate, Tryptophan, Tyrosine, Phenylalanine, Serine, Glycine, Cysteine</p>	[155]
<i>Arabidopsis thaliana</i>	Leaf	GC-MS	<p>Fumaric acid, Succinic acid, Fructose, Galactose, Raffinose, Galactinol, Maltitol, Glycine, Proline, Dehydroascorbic acid dimer, Hexadecanoic acid, Itaconic acid, Ethanolamine, Pyroglutamic acid, Xylose, Aspartic acid</p>	[156]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Arabidopsis thaliana</i>	Leaf rosettes	GC-MS	Galactose, Raffinose, Glucose, fructose, Galactinol, Serine, Threonic acid, Pyruvate, Citrate, Succinate, Malate, Proline, Glutamate, Glutamine, Melibiose, Tryptophan, Isoleucine, Alanine, Lysine, Aspartate, Tyrosine, Spermidine, Putrescine	[157]
<i>Arabidopsis thaliana</i>	Leaf	GC-TOF-MS	Galactinol, Raffinose, Proline, Allantoin, Tryptophan, Glyceric acid, Citrulline, Glutamine, Fructose, Arginine, Ornithine, Ascorbic acid, Asparagine, myo-Inositol, Fructose, Sucrose	[139]
<i>Miscanthus</i> species	Aerial organs	NMR	Leucine, Isoleucine, Valine, Lactic acid, Threonine, Alanine, Quinic acid, Proline, Glutamic acid, Glutamine, Shikimic acid, Malic acid, Asparagine, Glucose, Choline, Aconitic acid, Fructose, Raffinose, Trigonelline, Adenine, Phenylalanine, Tyrosine, Adenosine, Formic acid	[201]
<i>Solanum lycopersicum</i>	Pollen	LC-QTOF-MS	Spermidine, Spermine, Kaempferol dihexoside, Quercetin, Hydroxycinnamic, Tomatin, Hydroxytomatine	[158]
<i>Triticum aestivum</i>	Leaf	GC-MS	Fumaric acid, Pyruvic acid, Glyceric acid, Succinic acid, Asparagine, Alanine, Glutamic acid, Isoleucine, Phenylalanine, Proline, Fructose, Inositol, Xylitol, Mannitol, Hexadecanoic acid, Malic acid, Threonic acid, Glycine, Lysine, Serine, Melibiose	[159]
<i>Zea mays</i>	Leaf	NMR	Proline, Alanine, Choline, Inositol, Linoleyl fatty acid, Tyrosine, Isoleucine, Valine, Asparagine, Threonine, Aspartate, GABA, Fructose, Sucrose	[160]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
Metabolites under salt stress				
<i>Populus euphratica</i>	Leaf	GC-MS	Fumaric acid, Malic acid, Succinic acid, Glyceric acid, Fructose, Raffinose, Trehalose, Xylulose, Galactinol, Myo-inositol, Mannitol, Glycerol, β -alanine, Proline, Serine, Valine	[161]
<i>Vitis vinifera</i>	Shoot tips	GC-MS	Malate, Citrate, Isocitrate, Succinate, Fumarate, Chloride, Phosphate, Proline, Glutamate, Glutamine, Glycine, Serine, Asparagine, Fructose, Glucose, Aspartate	[162]
<i>Limonium latifolium</i>	Shoot and root	HPLC and NMR	Proline, Sucrose, Aspartate, β -alanine betaine, Fructose, Glucose, Glutamate, Choline-O-sulfate, Chiro-inositol, Myo-Inositol, Glutamine	[163]
<i>Zea mays</i>	Shoot and root	NMR	Alanine, Glutamate, Asparagine, GABA, Glycine betaine, Sucrose, trans-Aconitic acid, Malic acid, Succinic acid, Aspartate	[164]
<i>Thellungiella halophila</i> and <i>Arabidopsis thaliana</i>	Whole plant	GC-MS	Trehalose, Threonine, Sucrose, Succinic acid, proline, Raffinose, Serine, Phosphoric acid, Malic acid, Glutamic acid, Citric acid	[165]
<i>Arabidopsis thaliana</i>	Cell culture	LC-MS and GC-MS	Tryptophan, Fructose, Sucrose, Tyrosine, Phenylalanine, Shikimate, Pyruvate, Lactate, Malate, Fumarate, Succinate, Cysteine, Ethanolamine, Proline, Isoleucine, Leucine	[166]
<i>Oryza sativa</i>	Leaf	GC-MS	Pyruvic acid, Quinic acid, Gallic acid, L-Tyrosine, Shikimic acid, L-tryptophan, Serotonin, Kaempferol, Ferulic acid, Glucose-6-Phosphate, Lactobionic acid, Raffinose, D-Trehalose, Stearic acid, Palmitic acid, 4-Hydroxybenzoic acid, Vanillic acid	[167]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Oryza sativa</i>	Cultured cells	GC-MS	Succinate, Proline, GABA, Valine, Cysteine, Leucine, Asparagine, Alanine, Ornithine, Glucose, Fructose, Gallactose, Trehalose, Tyrosine, GABA, Methionine, serine, Phenylalanine	[168]
<i>Lotus japonicas</i>	Multiple tissues	GC/EI-TOF-MS	Proline, Leucine, Lysine, Serine, Threonine, Glycine, maltose, Glucose, Fructose, Pinitol, Arabitol, Glutamic, Succinic, Citric, Malic, Threonic	[169]
<i>Lotus creticus</i> , <i>Lotus corniculatus</i> , <i>Lotus tenuis</i>	Complete shoot	GC/EI-TOF-MS	Malic acid, Threonic acid, Ononitol, Erythronic acid, Succinic acid, Citric acid, Sucrose, Serine, Proline	[170]
<i>Hordeum vulgare</i>	Root and leaf	GC-MS	Proline, GABA, Putrescine, beta- Alanine, Aspartate, GABA, Glutamine, N-acetylglutamate, Proline, Phenylalanine, Putrescine, Serine, Threonine, Aconitate, Ascorbate, Monomethylphosphate, Isocitrate, 3PGA, 1,6-Anhydroglucose, 2-Keto-gluconate, 2-O-glycerol-b-D-galactose, Ribonate, Shikimate, Threonate-1,4-lactone, Fructose, Galactinol, Galacturonate, Gluconate, Glucose, Trehalose	[171]
<i>Hordeum vulgare</i> and <i>Hordeum spontaneum</i>	Root and leaf	GC-MS	Fumaric acid, Succinic acid, Isoleucine, Putrescine, Proline, asparagine, Leucine, Glycine, Serine, 3-PGA, Raffinose, Citric acid, Isocitric acid, Inositol, Trehalose, Mannitol, Sucrose	[172]
Metabolites under oxidative stress				
<i>Arabidopsis thaliana</i>	Cell suspension cultures	GC-MS	Alanine, Ascorbate, Asparagine, Aspartate, Gluconate, Malate, Ribose, Proline, Glycine	[173]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Arabidopsis thaliana</i>	Roots	GC-TOF-MS, LC-MS	GABA, O-acetyl serine (OAS), Pyruvate, Glucosinolates, Citrate, 2-oxoglutarate, Succinate, Fumarate, Malate, Alanine and Isoleucine	[174]
<i>Oryza sativa</i>	Cell cultures	CE-MS	Pyruvate, 3-phosphoglyceric acid, Dihydroxyacetone Phosphate, Fructose-6-phosphate, Glucose-1-phosphate (G1P), G6P, G3P, Phosphoenolpyruvate (glycolysis intermediates), O-Acetyl-L-serine, Cysteine, and γ -Glutamyl-L-cysteine	[175]
Metabolites under flooding stress				
<i>Glycine max</i>	Mitochondrial fractions	CE-MS	Citrate, Succinate aconitate, Gamma-amino butyrate (GABA), Pyruvate, NAD, NADH, UDP-glucose (UDPG), Glyceraldehyde-3-phosphate (GA3P), Glucose-1-phosphate (G1P) and 6-phospho-gluconate (6PG)	[176]
<i>Glycine max</i>	Root tips	CE-MS	Gamma-aminobutyric acid, Glycine, NADH ₂ , and Phosphoenol pyruvate	[177]
Metabolites under nutrient deficiency				
<i>Oryza sativa</i>	Leaf and root	GC-TOF-MS	Sugars and sugar phosphates, Tryptamine, Tyramine and 1,3-diaminopropane and Intermediates of TCA cycle	[178]
<i>Solanum lycopersicum</i>	Leaf	GC-MS	Cysteine, β -alanine, Glycine, Isoleucine, Phenylalanine, Valine, GABA, Lysin, Arabinose, Fucose, Inositol, Sucrose, Trehalose, 2-Oxoglutarate, Glutamate	[179]
<i>Arabidopsis thaliana</i>	Shoots	GC-MS	Malate, Fumarate, Ornithine, Raffinose, Trehalose, Succinate, GABA	[180]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Arabidopsis thaliana</i>	Seedlings	GC-MS, LC-MS	Flavonoids, Starch S-adenosylhomocysteine, O-acetylserine (OAS), Tryptophan, Glucosinolates, Uric acid	[181]
<i>Arabidopsis thaliana</i>	Seedlings	–	Sugar phosphates, Glycolysis, Glycerate-3- phosphate, Glycerate-2-phosphate, Phosphoenolpyruvate, Starch, Sucrose and Reducing sugars, Citrate, Fumarate, Malate, Oxoglutarate, Histidine, Arginine and threonine	[182]
<i>Phaseolus vulgaris</i>	Nodule	GC-TOF-MS	Putrescine, Picolinic acid, Malonic acid, Tartaric acid, Glyceric acid, Galactonic acid, Threitol, Sucrose, Glycine	[183]
<i>Phaseolus vulgaris</i>	Roots	GC-MS	Putrescine (agmatine), β-Alanine, 4-Aminobutyric acid, Threitol, Fructose, [926; Galactosyl glycerol (6TMS)]	[184]
<i>Hordeum vulgare</i> L.	Shoots and roots	GC-MS and LC-MS	Glucose-6-P, fructose-6-P, Inositol-1-P, Glycerol-3-P, 2-Oxoglutarate, Succinate, Fumarate and Malate	[185]
<i>Arabidopsis thaliana</i>	Shoots and roots	–	Sucrose, Fructose, Glucose, Malate, 2-Oxoglutarate [2-OG]	[186]
Metabolites under heavy metal stress				
<i>Brassica rapa</i>	Leaf and root	¹ H-NMR	Glucosinolates, Hydroxycinnamic acids, Amino acids, Phenolics	[187]
<i>Arabidopsis thaliana</i>	Seedlings	GC-TOF-MS	Alanine, β-Alanine, Proline, Serine, Putrescine, Sucrose GABA, Raffinose and Trehalose, α-Tocopherol, Campesterol, β-Sitosterol and Isoflavone	[188]
Metabolites under biotic stress				
<i>Oryza sativa</i>	Stem region, close to apical meristem	GC-MS	Margaric acid, Tetracosanoic acid, Arachidic acid, Galactose- 6-phosphate, Ribose, Inosine	[189]

(continued)

Table 2 (continued)

Plant species	Tissue of interest	Analytical approach	List of metabolites	References
<i>Oryza sativa</i>	Leaf	GC-TOF-MS, LC-TOF-MS	Acetophenone, 2-Phenylpropanol Xanthophylls, Alkaloids, Phenylalanine, Glutathione, Tyrosine	[190]
<i>Hordeum vulgare</i> L., <i>Oryza sativa</i> and <i>Brachypodium distachyon</i>	Leaf	GC-TOF-MS	Glutamate, Malate, Aspartate, GABA	[191]
<i>Triticum aestivum</i>	Spikelets	LC-ESI-LTQ-Orbitrap	Cinnamyl alcohol dehydrogenase, Caffeoyl-coa O-Methyltransferase, Caffeic acid Omethyltransferase, Flavonoid O-methyltransferase, Agmatine coumaroyl-transferase and Peroxidase	[192]

role of RFOs is not completely understood, there are several reports that show evidence for the strong correlation between accumulation of RFOs and development of desiccation tolerance [140].

Urano et al. [146] demonstrated metabolomic changes in *A. thaliana* (ecotype Col-0 (WT) and the *NCED3* knockout mutant) under drought stress. Many metabolites were identified including amino acids such as proline, raffinose family oligosaccharides, and γ -amino butyrate (GABA). In this study an *nc3-2* mutant that lacks the *NCED 3* gene is involved in the dehydration-inducible biosynthesis of abscisic acid (ABA), to determine the effect of ABA under drought stress. In combination with the transcriptome analysis, it was clearly demonstrated that ABA-dependent transcriptional regulation is important to activate metabolomic pathways like branched amino acids, polyamine and proline biosynthesis, GABA shunt, etc., but regulation of a raffinose biosynthetic pathway still remains unknown [146].

Polyamines such as putrescine, spermidine, and spermine are ubiquitous in nature and also provide protection to plants under drought stress [145, 197, 198]. GC-TOF-MS analysis by Do et al. [144] determined the level of selected metabolites related to polyamine metabolism in rice cultivar (*Oryza sativa* L. Ecotype indica and japonica) under moderate long-term drought stress. The combination of gene expression and GC-TOF-MS metabolite data showed coordinated adjustment of polyamine biosynthesis in order to facilitate the accumulation of spermine under drought stress conditions [144].

Gechev and collaborators combined transcriptomics and metabolomics to investigate desiccation tolerances in *Haberla rhodopensis* in four different conditions (i.e., well-watered, partially dehydrated, desiccated, and rehydrated) [141]. Transcripts of proteins involved in carbohydrate metabolism (e.g., genes that encode galactinol

synthase and stachyose synthase), sucrose synthase, and sucrose-6-phosphate synthase showed increased levels. These findings showed the importance of carbohydrate metabolism for protecting cells during desiccation. Genes that mainly encode proteins to prevent cellular damage and participate in antioxidant defense (e.g., LEA) were found to be most abundant in response to dehydration. In this study, the GC-TOF-MS approach was used to determine metabolite profiling combined with two different LC-MS approaches, which allowed the broad range of metabolome identification. This study also revealed that sucrose, maltose, and RFOs such as stachyose and verbascose accumulate in *H. rhodopensis* in significantly high levels upon dehydration. Additionally, other metabolites like amino acids, phenylalanine, and tyrosine were also observed in the dehydrated state, which suggests the activation of the shikimate pathway that results in the synthesis of antioxidants.

Skirycz and co-workers performed metabolite profiling of *Arabidopsis* leaves under mild drought stress. The response to stress in the growing and mature leaves was significantly distinct. Metabolites such as proline, erythritol, and putrescine were identified in mature leaves. Comparing the data with other studies revealed that decreases in the level of aspartate and increases in the level of proline are two common responses shared between mild and severely desiccated leaves [147, 199].

Metabolite profiling has been additionally carried out in other crop plants (maize, wheat, tomato, and soybean) under drought stress and it was observed that changes in the metabolite levels including branched chain amino acids are one of the common factors [143, 148, 149, 151].

A metabolic adjustment also depends on the severity of the stress. Maize was subjected to a drought stress for 17 days. GC-MS metabolomic analysis showed changes in the concentrations of 28 identified metabolites. Further, accumulation of carbohydrates, proline, amino acids, shikimate, serine, glycine, and aconitase were identified. Additionally, decreased levels of leaf starch, malate, fumarate, and 2-oxoglutarate were also observed in the drought-treatment course. However, between the 8th and 10th days, some metabolites were changed drastically, hence showing their dependency on stress severity [150].

The GC-MS metabolomic analysis was also performed on the moss *Physcomitrella patens* under drought stress. In this analysis, 2 weeks of physiological drought stress was applied, which showed that 26 metabolites were differentially affected in gametophores, including altrose, maltitol, L-proline, maltose, isomaltose, and butyric acid. More interestingly, a new compound, annotated as EITMS_N12C_ATHR_2988.6_1135EC44, was also accumulated specifically in response to drought stress in this moss [142].

5.2 Temperature Stress

A freezing environment leads to the formation of ice, which can seriously damage plant cells and cellular membranes. Many plant species develop freezing tolerance during their exposure to non-freezing low temperatures, and this process is known as

“cold acclimatization” [200]. The molecular basis of this process has been widely studied. The first metabolomic studies of cold acclimatization were performed by two groups in 2004. Cook et al. [152] compared the metabolomic changes during cold acclimatization in *A. thaliana* (ecotype Wassilewskija – 2 (Ws-2) and Cape Verde islands-1 (CVi-1)). A metabolome of the Ws-2 plant was significantly changed in response to low-temperature stress. Seventy-five percent of metabolites monitored were found to increase under cold-acclimated plants, which include amino acids such as proline and sugars (glucose, fructose, inositol, galactinol, raffinose, and sucrose). Additional changes were also identified with increased levels of trehalose, ascorbate, putrescine, citrulline, and some TCA-cycle intermediates. However, there were considerable overlaps in the metabolite changes between the two ecotypes in response to low temperature [152].

Time-course metabolomic analysis (from cold to heat conditions) showed an increase in the pool size of amino acids derived from pyruvate, oxaloacetate, polyamine precursors, and other compatible solutes [154]. The study concluded that the majority of the heat shock metabolite responses were shared with cold stress, while heat shock had a less pronounced effect on metabolism. Transcriptomics analysis revealed the regulation of GABA shunt and the accumulation of proline under a cold condition that was obtained by transcriptional and post-transcriptional processes [155]. Espinoza et al. [153] studied the effect of diurnal gene/metabolite regulation during cold acclimatization by using metabolomics and transcriptomics. Approximately 30 % of identified/analyzed metabolites showed the circadian rhythm in their pool size and low temperatures affected the cyclical pattern of metabolite abundance [153].

The number of important traits in plants, such as stress resistance, post-harvesting etc., largely dependents upon the metabolic content, which can be used for the manipulation of the metabolic phenotype via a classical breeding method. A study performed by Korn et al. [156] combined GC-MS metabolite profiling and statistical methods to decode or identify the combination of metabolites that can predict the freezing tolerance in *Arabidopsis*. One of the identified candidates was raffinose, which can also be considered as a good marker for freezing tolerance [156].

Metabolomics was also used for functional characterization of candidate genes involved in cold acclimatization. One of the well-characterized genes is a C-repeat binding factor (CBF) that increases freezing tolerance by multiple mechanisms. Cook et al. [152] investigated the regulation of CBF and its effect on the metabolome of the plant under low temperatures using GC-MS analysis. Metabolite profiling of the non-acclimated plants that over-expressed the CBF 3 was similar to that of the cold acclimated Ws-2 ecotype. Hence these data indicate that the CBF pathway plays a prominent role in determining metabolite regulation at low temperatures. Further, the analysis reveals the accumulation of raffinose and galactinol, which are synthesized through the action of CBF-targeted genes *AtGolS3*.

In another study, Wienkoop et al. investigated the dynamics of metabolite-protein covariance networks and the relation of starch metabolism during cold acclimation [139]. This study revealed that raffinose accumulation belongs to general cold and

heat temperature stress responses and that starch metabolism can be compensated by increased sucrose synthesis for cold adaptation processes. A follow-up study revealed the essential role of starch metabolism during cold adaptation in different *Arabidopsis* ecotypes, thus demonstrating that different ecotypes developed different biochemical strategies to cope with cold acclimation [157]. In another study, Doerfler et al. investigated the interface of primary and secondary metabolism as a response to cold and light stress [81, 82]. *Arabidopsis* accumulated huge amounts of flavonoid structures due to the combined cold and light stress and a novel algorithm for metabolite identification in non-targeted LC-MS metabolomics data was able to identify new potential flavonoid structures as anti-oxidative response factors [82]. Several other metabolomics analyses have been performed in crop and grass species (tomato, wheat, maize, and miscanthus) [158–160, 201] that determined the diverse range of metabolites in plants to confers stress tolerance (Table 2).

5.3 Salt Stress

Increasing salt concentration in soil damages plants in various ways: (1) it hampers the uptake of water and nutrients from the environment, which in turn reduces the water potential of the soil and leads to osmotic stress. Salt stress reduces plant growth and damages cells/tissues. (2) The steady accumulation of sodium ions in plant tissues inhibits essential cellular processes [202, 203]. Plants have adapted strategies to cope with salt stress, which includes adjustment of metabolic status [204].

Gong et al. [165] conducted metabolite profiling on *Thellungiella halophila*, a distant relative of *A. thaliana*, under salt stress. This plant shows “extremophile” characteristics manifested by extreme tolerance to a variety of abiotic stresses like low humidity, freezing, and high salinity (it can even grow and reproduce in a 500 nM NaCl concentration) [205]. A comparative metabolomics study between *T. halophila* and *Arabidopsis* (under controlled conditions) showed increased levels of proline in both the species along with inositols, hexoses, and complex sugars. The concentration of these metabolites was higher in *T. halophila*. Transcriptome analysis showed similar results suggesting that *T. halophila* is primed for acclimatization under stress conditions [165]. Kim et al. [166] showed the effect of high salt concentrations on primary metabolism in a cell culture of *A. thaliana* (T87 cultured cells). The obtained results showed that the methylation cycle and phenylpropanoid pathways are synergistically induced over the short term in response to salt stress. Long-term responses include co-induction of glycolysis and sucrose metabolism as well as co-reduction of the methylation cycle [166].

There have been several metabolomic studies performed to assess the metabolic effect of salinity in various crops and other plant species, which include tomato [206, 207], grapevine [162], poplar [161], sea lavender (*Limonium latifolium*), [163] and rice [167]. A study performed by Sanchez and co-workers extensively used the

integrated approach of genomic, transcriptomic, and metabolomic analysis on *L. japonicus* and other lotus species in long-term regimes of non-lethal salt stress. The metabolomic changes were characterized by steady-state increased levels of amino acids, sugars, and polyols with decreases in most organic acids [169]. A comparative metabolomic study between extremophile (*Lotus crticus*) and glycophyte (*Lotus corniculatus*) generated similar metabolomic responses under salt stress conditions [170].

Barley (*Hordeum vulgare*) cultivars that differ in their salt tolerance capacity were subjected to metabolite profiling. In this study long-term salt stress was applied to the barley plants [171]. The more tolerant cultivar (Sahara) demonstrated increased levels of hexose phosphates and intermediates of the TCA cycle, which include citrate, aconitate, isocitrate, α -ketoglutarate, succinate, and malate. The levels of metabolites remain unchanged in the less tolerant cultivar (Clipper) during salt stress. From this study it was proposed that accumulation of proline, γ -amino butyric acid (GABA), and other amino acids in Clipper showed growth or induction of leaf necrosis under salt stress; however, accumulation of these metabolites does not indicate the phenomenon of acclimatization [171]. Another comparative study of barley (cultivated vs. wild) demonstrated that the wild type of barley is more salt tolerant than cultivated barley and possesses an improved ability to regulate osmotic stress through the accumulation of more carbohydrates and proline in its roots. GC-MS-based metabolite profiling led to the identification of 82 metabolites, which include water-soluble carbohydrates (sucrose, trehalose, and raffinose) and proline, which were predominant and contribute potentially to salt tolerance in roots [172]. A change in the amino acid metabolism in leaves seems critical to develop a salt tolerance mechanism.

Rice is one of the most sensitive crops to elevated salt concentration because its roots are highly permeable to sodium ions present in the soil [203]. A GC-TOF-MS metabolomic analysis revealed the lower levels of TCA cycle intermediates and organic acid in the roots of tolerant rice cultivars in comparison to the sensitive cultivar. In addition, accumulation of amino acids was also observed in the tolerant cultivar [167]. Liu et al. [168] conducted metabolite profiling of rice cell culture focusing on the initial phase of salt stress that was exclusively characterized by osmotic stress. Further, glucose, fructose, galactose, hexose phosphates, glucose-6-phosphate, and fructose-6-phosphate accumulated in rice suspension culture subjected to 100 mM NaCl for 1–24 h. Several studies on different rice cultivars showed a decrease in the TCA cycle-dependent organic acids and accumulation of various amino acids [168].

Maize plants exposed to salt stress (50–150 mM NaCl saline solution) showed an increase in the levels of sucrose and alanine, but the levels of glucose were decreased in roots and shoots. Other osmoprotectants like GABA, malic acid, and succinate showed increased levels in roots, whereas acetoacetate showed decreased levels. Simultaneously glutamate, asparagine, and glycine betaine showed increased levels in shoots, whereas malic acid and trans-aconitic acid showed decreased levels. Increased metabolic response was more evident in shoots than in roots [164].

5.4 Oxidative Stress

Oxidative stress is one of the major limiting factors for plant growth. It occurs by overproduction of ROS, for example hydrogen peroxide (H_2O_2), superoxide ($\text{O}_2^{\cdot-}$), and singlet oxygen ($^1\text{O}_2$). A wide range of abiotic stresses like high light, low temperature, drought, and salt stress can cause oxidative stress [208]. ROS are produced constantly in the cell as a by-product of aerobic metabolism even under non-stress conditions, particularly during photosynthesis, in mitochondria via the mitochondrial electron transport chain, and in peroxisomes upon photorespiration and β -oxidation of fatty acids. However, the cellular antioxidant system can detoxify ROS via the ascorbate glutathione (GSH) cycle; a balanced cellular redox-status is maintained. Although low levels of ROS are important for signaling and response for certain stress signals, elevated ROS levels can contribute to a plant's defense program by initiating programmed cell death, especially during pathogen attack [209].

In a study performed by Baxter et al. [173], heterotrophic *Arabidopsis* cells were treated with menadione, which enhances ROS production via the electron transport chain and hence changes metabolite abundance. Metabolomic abundance was quantified using ^{13}C -labelling kinetics. It was observed that sugar phosphates related to glycolysis and oxidative pentose phosphate pathways (OPPPs) were accumulated, which directs the flow of the glycolytic carbon into the OPPP to provide NADPH for antioxidant activity. In addition, levels of ascorbate decreased and the accumulation of its degraded products like threonate was observed, which indicated activation of the antioxidative pathway in menadione-treated cells. Reduction in glycolytic activity probably leads to decreased levels of amino acids derived from glycolytic intermediates. Further inhibition of the TCA cycle intermediates was also observed and confirmed by ^{13}C redistribution analysis [173].

Lehmann et al. [174] also performed a similar kind of study considering metabolite profiling and ^{13}C -redistribution analysis on menadione-treated *Arabidopsis* roots. The results obtained were distinct from the heterotrophic cell study. In this analysis, roots showed pronounced accumulation of GABA, O-acetylserine (OAS), pyruvate, glucosinolates, and other amino acids. It is likely that cellular oxidation inhibits sulfur assimilation and leads to the accumulation of OAS [174]. Similarly, rice cell cultures treated with menadione redirected the carbon flux from glycolysis through the OPP pathway and subsequently increased the levels of NADPH. CE-MS analysis of these rice cultures showed depletion of sugar phosphates like pyruvate, 3-phosphoglyceric acid, dihydroxyacetone phosphate, fructose-6-phosphate, glucose-1-phosphate (G1P), G6P, G3P, phosphoenolpyruvate (glycolysis intermediates), and TCA intermediates like 2-oxoglutarate, aconitate, citrate, fumarate, isocitrate, malate, and succinate; followed by the increase in OPP pathway intermediates like 6-phosphogluconate, ribose 5-phosphate, and ribulose 5-phosphate. Additionally, an increase in the biosynthesis of GSH and its intermediates (O-acetyl-L-serine, cysteine, and γ -glutamyl-L-cysteine) was also observed in the menadione-treated rice cell cultures [175]. Down-regulated expression of manganese superoxide

dismutase (MnSOD) levels, which cause oxidative stress on the metabolome, was also observed. GC-MS metabolite profiling in *Arabidopsis* revealed a redox shift in mitochondria, followed by a specific decrease in TCA cycle flux, due to the inhibition of aconitase and isocitrate dehydrogenase [210].

5.5 *Flooding*

Flooding imposes severe stress on plants. This is principally because excess water in the surroundings can deprive them of oxygen and CO₂, which in turn hampers the process of photosynthesis and reduces growth and grain yield [211]. Mitochondrial fractions from the roots and hypocotyls of 4-day-old soybean seedlings that had been in flooding stress for 2 days were subjected to proteomics and metabolomics analysis. Proteins and metabolites related to the TCA cycle (like citrate, succinate, and aconitate), GABA shunt, and amounts of NADH and NAD were up-regulated under stress conditions, but ATP was significantly decreased by flooding stress [176]. Similarly, Komatsu et al. [177] investigated the root tips of soybean under flooding stress – in total 73 flood responsive metabolites were identified using capillary electrophoresis-mass spectrometry. The levels of gamma-aminobutyric acid, glycine, NADH₂, and phosphoenol pyruvate were increased under flooding stress conditions [177].

5.6 *Nutrient Deficiency*

Nutrients are essential for plant growth and development. They can affect and regulate fundamental processes of plant physiology like photosynthesis and respiration. Depending upon the growth requirement of the plants, nutrients are referred to as either macronutrients (e.g., nitrogen, phosphorus, potassium, sulfur, and magnesium) or micronutrients (e.g., iron, zinc, etc.). Nutrient starvation or limitation of macronutrients has direct effects on the metabolism of the plant since most organic molecules are made up of a combination of these elements.

5.6.1 **Nitrogen (N)**

Nitrogen is the most important nutrient required by plants and its metabolism is highly coordinated with carbon metabolism in the fundamental process of plant growth. Paddy fields of rice plants preferentially use ammonium as a source of inorganic nitrogen. The conversion of ammonium to glutamine is catalyzed by glutamine synthetase (GS). Kusano et al. [178] performed comparative metabolomic analysis between the rice mutant lacking the OsGS1;1 gene and the wild type. The results revealed that mutant lines showed retardation in shoot growth in the presence

of ammonium compared to the wild type. Overaccumulation of free ammonium in the leaf sheath and roots of the mutant lines demonstrated the importance of the *OsGs1;1* gene for ammonium assimilation. The metabolomic profile of the mutant line revealed decreased levels of sugars, amino acids, and intermediates of the TCA cycle. In contrast, overaccumulation of secondary metabolites was observed particularly in roots under a continuous supply of ammonium [178]. Further, the effect of nitrogen deficiency at the metabolite levels in tomato leaves was investigated by Urbanczyk-Wochniak and Fernie [179]. Based on the analysis, the decreased levels of 2-oxoglutarate, as well as other TCA-cycle intermediates including citrate, isocitrate, succinate, fumarate, and malate, were observed under nitrogen stress [179]. Similarly, Tschoep et al. [180] analyzed the effect of mild nitrogen limitation in *Arabidopsis*, where the levels of malate and fumarate showed significant decreased levels [180]. These findings were in agreement with a previous study performed in tomato leaves [179].

5.6.2 Sulfur (S)

Sulphur is another macronutrient essential for the synthesis of sulfur-containing amino acids like cysteine and methionine as well as a wide range of sulfur-containing metabolites like glutathione. Sulfur stress has been well studied using metabolomics by several groups, and has been well elaborated by Hoefgen and Nikiforova [212]. Nikiforova et al. used GC-MS and LC-MS profiling methods to monitor the response of 134 metabolites and 6,023 unknown peaks of non-redundant ion traces under sulfur stress. Based on the profiling data, the coordinated network of metabolic regulation was successfully reconstructed under sulfur stress [213]. These data were also analyzed together with transcriptomic data in order to generate a gene-metabolite correlation network in *Arabidopsis* under sulfur stress [181].

5.6.3 Phosphorus (P)

Morcuende et al. [182] analyzed *Arabidopsis* seedlings grown in liquid culture under phosphorus starvation. The metabolite profile revealed the levels of sugar phosphates were decreased but other metabolites like glycolysis, glycerate-3-phosphate, glycerate-2-phosphate, and phosphoenolpyruvate were increased in P-deficient seedlings. P-deficient seedlings also showed accumulation of starch, sucrose, and reduced sugars as well as a general increase of organic acids including citrate, fumarate, malate, and oxoglutarate. The levels of aromatic amino acids like histidine, arginine, and threonine did not alter or increased slightly. Together with transcriptomic data, analysis of metabolites revealed that phosphorus deprivation leads to a shift towards the accumulation of carbohydrates, organic acids, and amino acids [182]. Hernández et al. performed metabolite profiling to understand the effect of phosphorus deficiency in the roots [184] and nodules [183] of the common bean.

Huang et al. [185] performed metabolomic analysis in both shoots and roots of phosphorus-deficient barley. Severe phosphorus deficiency decreased the levels of phosphorylated intermediates (glucose-6-P, fructose-6-P, inositol-1-P, and glycerol-3-P) and organic acids (2-oxoglutarate, succinate, fumarate, and malate). It was also identified that phosphorus-deficient plants reconstruct carbohydrate metabolism initially in order to reduce phosphorus consumption, which consequently reduces the levels of organic acid in the TCA cycle [185].

5.6.4 Potassium (K)

Potassium (K) plays essential roles as a major cation in plants and as a cofactor of enzymes. Armengaud et al. [186] performed metabolite profiling in order to identify metabolic targets of potassium stress. Metabolite profiles of *Arabidopsis* plants (roots and shoots) under low-K concentration revealed increases in the levels of soluble sugars (sucrose, fructose, and glucose) and a slight net increase of total protein content and the overall amino acid level. Additionally, a strong decrease in pyruvate and organic acids was observed only in the roots but not in the shoots [186].

5.6.5 Heavy Metals

Heavy metals such as cadmium (Cd), cesium (Cs), lead (Pb), zinc (Zn), nickel (Ni), and chromium (Cr) are major soil pollutants (from the beginning of the industrial revolution) causing stress on plants. Heavy metals induce enzyme inhibition, cellular oxidation, and metabolic perturbation, resulting in growth retardation and plant death in extreme instances [214]. Sometimes essential nutrients including copper (Cu), iron (Fe), and manganese (Mn) can cause heavy metal stress if present in an inappropriate concentration. Jahangir et al. [187] analyzed the effects of Cu, Fe, and Mn on the metabolite levels of *Brassica rapa* (leaves and roots). The levels of metabolites such as glucosinolates and hydroxycinnamic acids as well as primary metabolites such as carbohydrates and amino acids were affected constitutively [187]. Similarly, *Arabidopsis* plants treated with Cd showed increased levels of alanine, β -alanine, proline, serine, putrescine, sucrose, and other metabolites with compatible solute-like properties, notably GABA, raffinose, and trehalose. This study also indicated a significant increase in the concentrations of α -tocopherol, campesterol, β -sitosterol, and isoflavone. When taken together these data indicate an important role of antioxidant defense in the mechanisms of resistance to cadmium stress [188].

5.7 Biotic Stress

In order to combat the stress from pathogens and pests, plants use complex chemical machinery as a major defense mechanism. The metabolic response of plants to biotic stresses depends highly on tissues, species, and plant-pathogen or pest interactions. A number of metabolites have been identified as metabolic biomarkers for biotic stress in diverse plant species [215]. For example, metabolic profiling of host tissues from the three rice varieties TN1, Kavya, and RP2068 exposed to gall midge biotype 1 (GMB1) was performed using GC-MS. In total, 16 fatty acids (such as unsaturated linoleic acid) together with two amino acids (glutamine and phenylalanine) were identified as the major components of the resistance features of gall midge (*Orseolia oryzae* Wood-Mason)-resistant rice varieties [189]. Similarly, metabolite profiling of sensitive and tolerant rice cultivars subjected to BLB (bacterial leaf blight) caused by *Xanthomonas oryzae* pv. *oryzae* (Xoo) led to the identification of several specific metabolites, such as acetophenone, xanthophylls, alkaloids, carbohydrates, and lipids [190]. Metabolomic analysis of barley, rice, and purple false brome grass, revealed identical changes in the metabolic patterns in which malate, polyamines, quinate, and non-polymerized lignin precursors accumulate during infection by *Magnaporthe oryzae* [191]. Phenylpropanoids are the precursors of lignin and constitute an important component of the plant stress defense mechanism by modulating cell wall composition and stiffness in roots. The thickened cell wall may help to defend against pathogen infection in the plant. Accumulation of phenylpropanoid and phenolic compounds was reported in response to *Fusarium graminearum* in wheat [192].

5.8 Stress Combination

In the natural habitat, plants are actually subjected to combinations of abiotic stress. Some of these stress conditions are already in combination, for example, high salt concentration leads to ionic and osmotic stresses. Although the metabolic response of plants under single stress conditions has been studied extensively as detailed in the previous section [216], Rizhsky et al. [217] applied the combination of drought and heat stress in *Arabidopsis*. The metabolite profiling showed high-level accumulation of sucrose and other sugars. Interestingly, proline was not accumulated in stress combination and hence it was concluded that sucrose replaces proline as an osmoprotectant in plants in order to avoid combined stress conditions because proline shows high-level toxic effects under heat stress conditions [217]. Wulff-Zottele et al. [218] also analyzed the effect of high light irradiance and sulfur depletion. In this study, it was observed that proline accumulated in a differential time course and other metabolites like raffinose and putrescine replaced proline during its delayed accumulation under stress [218].

6 Metabolite Accumulation: Adjustment in Response to Stress Conditions

Under abiotic stress conditions, a common defensive mechanism is activated in plants that leads to the production and accumulation of compatible solutes (metabolites). These are small molecular weight osmoprotectants that are diverse in nature, and include amino acids like asparagine, proline, and serine, and amines like polyamines, glycine betaine, and GABA (γ -amino-N-butyric acid). Additionally, they also include carbohydrates (e.g., fructose, sucrose, and trehalose), raffinose, and polyols (myo-inositol, D-pinitol), and antioxidants such as glutathione (GSH) and ascorbate [219]. These osmolytes have high levels of solubility/permeability in the cell and lack enzyme inhibition activities even at high concentrations. Accumulation of the solutes in response to stress is not only observed in plants, it is also a defense mechanism triggered by animal cells, bacteria, and marine algae, which indicate an evolutionarily conserved trait [196, 220, 221].

Several studies have been performed to understand the beneficial effects of these metabolites in plant tolerance against environmental stimuli. Proline is one of the best examples, and the correlation between proline accumulation and stress tolerance has been well described in Bermuda grass during water stress conditions [222]. In the follow-up studies, several research works were conducted to prove and correlate the importance of proline accumulation in plants under stress conditions. Based on the findings it can be concluded that proline can act as a protein stabilizer, a signaling molecule, a ROS scavenger (against several abiotic stresses), and a cryoprotectant/osmoprotectant in plants under stress conditions [223, 224]. Figure 3 represents the biosynthesis of the key metabolites that accumulate in plants in response to the stress condition.

Proline metabolism and its regulations are well characterized in plants (Fig. 3(1)). Proline is synthesized in cytoplasm or chloroplast from glutamate, which reduces to glutamate semialdehyde (GSA) by Δ -1-pyrroline-5-carboxylate synthetase (P5CS). GSA can spontaneously convert into pyrroline-5-carboxylate (P5C), which is further reduced by P5C reductase (P5CR) to proline. Proline is degraded within the mitochondrial matrix by proline dehydrogenase (ProDH) and P5C dehydrogenase (P5CDH) to glutamate [225]. Under stress conditions proline synthesis is stimulated, whereas during recovery from stress catabolism of proline is enhanced. In tobacco and petunia, over-expression of P5CS led to increasing accumulation of proline and showed enhanced tolerance to salt and drought [226]. *Arabidopsis* P5CS1 knock-out plants were impaired in proline synthesis and showed hypersensitivity to salt stress [227]. ProDH antisense *Arabidopsis* accumulated more proline, which shows enhanced tolerance against low temperature and high salinity [228]. An alternative pathway, mitochondrial P5C, can be produced by δ -aminotransferase (δ -OAT) from ornithine. Over-expression of *Arabidopsis* δ -OAT enhances proline levels, which in turn increases stress tolerance in rice and tobacco [229]. The core enzymes P5CS, P5C, P5CR, ProDH, and OAT are responsible for maintaining the balance between biosynthesis and catabolism of proline.

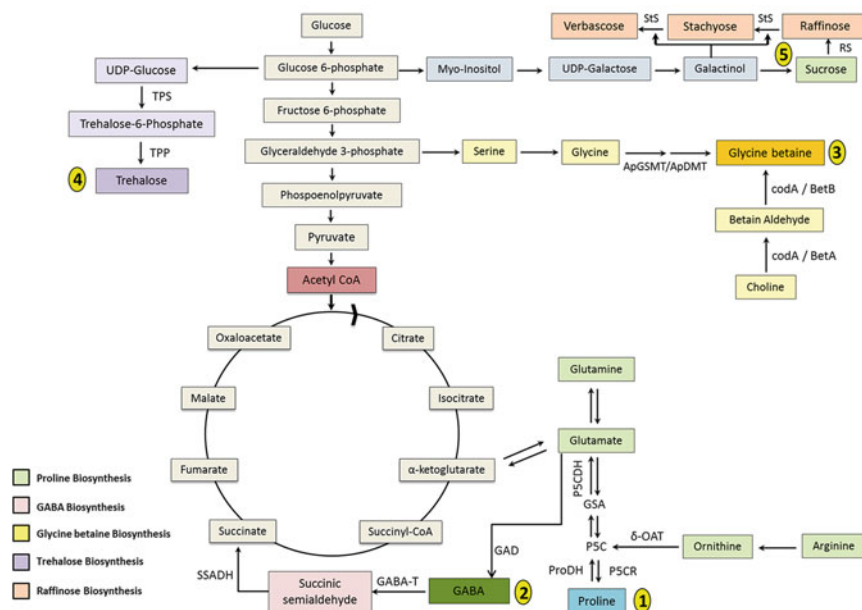


Fig. 3 Biosynthetic pathway of the metabolites that are accumulated under stress conditions (e.g. drought, temperature, salt, etc.). (1) proline, (2) GABA, (3) glycine betaine, (4) trehalose, (5) raffinose

Extensive research work has been performed for other metabolites such as γ -aminobutyric acid (GABA), glycine betaine, trehalose, raffinose, and polyamines (like putrescine, spermidine, and spermine) etc.; these metabolites have also been proven to be efficient protectors against some abiotic stresses. GABA is a non-protein amino acid that accumulates rapidly under high-stress levels [230–232]. GABA is mainly synthesized from glutamate in cytosol and then transported into mitochondria. GABA-T (GABA transaminase) and SSADH (succinic semialdehyde dehydrogenase) convert GABA into succinate, which enters into the TCA cycle (Fig. 3(2)) [233]. GABA metabolism plays a major role in the carbon-nitrogen balance and ROS scavenging activity [234, 235]. GABA shunt also plays an important role in stress tolerance. Enzymes involved in GABA metabolism showed enhanced enzyme activity under salt stress [232]. GABA-T deficient *Arabidopsis* mutants showed hypersensitivity against ionic stress – increased levels of amino acids were accumulated while carbohydrate levels were diminished [232]. Similarly, disruption of the SSADH gene showed ROS accumulation and hypersensitivity against UV-B and heat stress [236].

Glycine betaine (GB) is a quaternary ammonium compound that occurs widely in the plant kingdom [237]. GB is synthesized from choline and glycine [238]. *Arabidopsis* and many other crops species do not accumulate GB. Plants with natural production of GB under stress conditions (cold, drought, and salt) showed enhanced accumulation of GB (Fig. 3(3)) [238]. In salt-tolerant species, a

significant level of GB accumulation was determined. GB protects photosystem II, stabilizes the membrane, acts as a molecular chaperone, maintains water balance, and provides protection from oxidative stress [239–241].

Trehalose is a non-reducing disaccharide that accumulates in higher concentrations under some desiccation-tolerant plants like *Myrothamnus flabellifolius* [242]. At high levels, trehalose can act as an osmolyte to stabilize proteins [243]. In plants, trehalose is present in a small amount, but under high-stress conditions trehalose increases moderately [154, 244]. Biosynthesis of trehalose takes place in two steps: trehalose-6-phosphate synthase (TPS) produces trehalose-6-phosphate (T6P) from UDP-glucose and glucose-6-phosphate followed by dephosphorylation of trehalose by trehalose-6-phosphate phosphatase (TPP) [243] (Fig. 3(4)). Several research studies demonstrated the role of trehalose in stress tolerance; transgenic expression of the genes involved in trehalose biosynthesis enhances the tolerance of plants against abiotic stress. Heterologous expression of genes involved in the trehalose pathway from *E. coli* or *Saccharomyces cerevisiae* showed enhanced tolerance to abiotic stresses in several plants [245]. Further, over-expression of the TPS isoform conferred enhanced resistance in rice against salt, cold, and drought stress [246]. Loss of TPS5 (TPS with TPP domain) function lowered the basal thermotolerance in *Arabidopsis* [247]. Panikulangara et al. [248] demonstrated that levels of trehalose and activity of TPP increase under cold stress in rice, and overexpression OsTPP1 in rice showed enhanced tolerance against cold and salinity, although trehalose content was not observed to be increased [248].

Raffinose family oligosaccharides (RFOs) include raffinose, stachyose, and verbascose, which accumulate in various plant species (in desiccated seeds and leaves) under environmental stress like cold, heat, drought, or salinity [249]. RFO biosynthesis is initiated by formation of galactinol from *myo*-inositol and UDP-galactose by galactinol synthase (Gols). Sequential addition of galactose units by galactinol to sucrose leads to the formation of raffinose, and a higher class of RFO (Fig. 3(5)) [249]. The complete biosynthetic pathway of RFOs under stress conditions is not completely known. Research work performed by Taji et al. [250] and Nishizawa et al. [251] demonstrated that over-expression of *Arabidopsis* Gols1 and Gols2 accumulates high levels of galactinol and sucrose in *Arabidopsis*, which showed enhanced tolerance under drought and salt stress [250, 251].

Metabolomic adjustments play a very important role in plant survival; therefore, regulation of the metabolic pathways (biosynthesis and catabolism) is very critical in order to improve tolerance mechanisms against environmental stimuli in plants.

7 Conclusion and Outlook

Metabolomics is the comprehensive platform that can identify and quantify small molecules present in plants that lead to the ultimate expression of its genotype in response to environmental stimuli (abiotic and biotic stresses). Information obtained from the metabolomic studies in plants against different abiotic and biotic stresses

have provided relevant information about the specific metabolites that are directly involved in physiological and biochemical changes. In an applied context, metabolomic approaches provide a broader, deeper, and integral perspective of the metabolic profiles in the acclimatization of plants against environmental stress. The obtained information is also transferable to the sensitive and economically important crops, to improve their adaptation strategies towards adverse conditions. The application of more advanced metabolomics tools will accelerate and improve plant breeding approaches, which will surely lead to the next generation of crops that are more tolerant to abiotic and biotic stresses worldwide.

References

1. Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, Town CD, Venter JC, Fraser CM, Tabata S, Nakamura Y, Kaneko T, Sato S, Asamizu E, Kato T, Kotani H, Sasamoto S, Ecker JR, Theologis A, Federspiel NA, Palm CJ, Osborne BI, Shinn P, Conway AB, Vysotskaia VS, Dewar K, Conn L, Lenz CA, Kim CJ, Hansen NF, Liu SX, Buehler E, Altafi H, Sakano H, Dunn P, Lam B, Pham PK, Chao Q, Nguyen M, Yu GX, Chen HM, Southwick A, Lee JM, Miranda M, Toriumi MJ, Davis RW, Wambutt R, Murphy G, Dusterhoft A, Stiekema W, Pohl T, Entian KD, Terryn N, Volckaert G, Salanoubat M, Choisne N, Rieger M, Ansoerge W, Unseld M, Fartmann B, Valle G, Artiguenave F, Weissenbach J, Quetier F, Wilson RK, De La Bastide M, Sekhon M, Huang E, Spiegel L, Gnoj L, Pepin K, Murray J, Johnson D, Habermann K, Dedhia N, Parnell L, Preston R, Hillier L, Chen E, Marra M, Martienssen R, McCombie WR, Mayer K, White O, Bevan M, Lemcke K, Creasy TH, Bielke C, Haas B, Haase D, Maiti R, Rudd S, Peterson J, Schoof H, Frishman D, Morgenstern B et al (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
2. Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchinson D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong JP, Miguel T, Paszkowski U, Zhang SP, Colbert M, Sun WL, Chen LL, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu YS, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* 296:92–100
3. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, Cao ML, Liu J, Sun JD, Tang JB, Chen YJ, Huang XB, Lin W, Ye C, Tong W, Cong LJ, Geng JN, Han YJ, Li L, Li W, Hu GQ, Huang XG, Li WJ, Li J, Liu ZW, Li L, Liu JP, Qi QH, Liu JS, Li L, Li T, Wang XG, Lu H, Wu TT, Zhu M, Ni PX, Han H, Dong W, Ren XY, Feng XL, Cui P, Li XR, Wang H, Xu X, Zhai WX, Xu Z, Zhang JS, He SJ, Zhang JG, Xu JC, Zhang KL, Zheng XW, Dong JH, Zeng WY, Tao L, Ye J, Tan J, Ren XD, Chen XW, He J, Liu DF, Tian W, Tian CG, Xia HG, Bao QY, Li G, Gao H, Cao T, Wang J, Zhao WM, Li P, Chen W, Wang XD, Zhang Y, Hu JF, Wang J, Liu S, Yang J, Zhang GY, Xiong YQ, Li ZJ, Mao L, Zhou CS, Zhu Z, Chen RS, Hao BL, Zheng WM, Chen SY, Guo W, Li GJ, Liu SQ, Tao M, Wang J, Zhu LH, Yuan LP, Yang HM (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science* 296:79–92
4. Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, Egholm M, Knight J, Bogden R, Li CB, Shuang Y, Xu X, Pan SK, Cheng SF, Liu X, Ren YY, Wang J, Albiero A, Dal Pero F, Todesco S, Van Eck J, Buels RM, Bombarely A, Gosselin JR, Huang MY, Leto JA, Menda N, Strickler S, Mao LY, Gao S, Teclé

- IY, York T, Zheng Y, Vrebalov JT, Lee J, Zhong SL, Mueller LA, Stiekema WJ, Ribeca P, Alioto T, Yang WC, Huang SW, Du YC, Zhang ZH, Gao JC, Guo YM, Wang XX, Li Y, He J, Li CY, Cheng ZK, Zuo JR, Ren JF, Zhao JH, Yan LH, Jiang HL, Wang B, Li HS, Li ZJ, Fu FY, Chen BT, Han B, Feng Q, Fan DL, Wang Y, Ling HQ, Xue YBA, Ware D, McCombie WR, Lippman ZB, Chia JM, Jiang K, Pasternak S, Gelley L, Kramer M, Anderson LK, Chang SB, Royer SM, Shearer LA, Stack SM, Rose JKC, Xu YM, Eannetta N, Matas AJ, McQuinn R, Tanksley SD, Camara F, Guigo R, Rombauts S, Fawcett J, Van De Peer Y, Zamir D, Liang CB, Spannagl M, Gundlach H, Bruggmann R et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–641
5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, Mckusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei YD, Li ZY, Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun JT, Wang ZY, Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH et al (2001) The sequence of the human genome. *Science* 291:1304
 6. Kehoe DM, Villand P, Somerville S (1999) DNA microarrays for studies of higher plants and other photosynthetic organisms. *Trends Plant Sci* 4:38–41
 7. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene-expression. *Science* 270:484–487
 8. Chaturvedi P, Doerfler H, Jegadeesan S, Ghatak A, Pressman E, Castillejo MA, Wienkoop S, Egelhofer V, Firon N, Weckwerth W (2015) Heat-treatment-responsive proteins in different developmental stages of tomato pollen detected by targeted mass accuracy precursor alignment (tMAPA). *J Proteome Res* 14:4463–4471
 9. Chaturvedi P, Ghatak A, Weckwerth W (2016) Pollen proteomics: from stress physiology to developmental priming. *Plant Reprod* 29:119–132
 10. Chaturvedi P, Ischebeck T, Egelhofer V, Lichtscheidl I, Weckwerth W (2013) Cell-specific analysis of the tomato pollen proteome from pollen mother cell to mature pollen provides evidence for developmental priming. *J Proteome Res* 12:4892–4903
 11. Ghatak A, Chaturvedi P, Nagler M, Roustan V, Lyon D, Bachmann G, Postl W, Schrofl A, Desai N, Varshney RK, Weckwerth W (2016) Comprehensive tissue-specific proteome analysis of drought stress responses in *Pennisetum glaucum* (L.) R. Br. (Pearl millet). *J Proteome* 143:122–135
 12. Ghatak A, Chaturvedi P, Paul P, Agrawal GK, Rakwal R, Kim ST, Weckwerth W, Gupta R (2017) Proteomics survey of Solanaceae family: current status and challenges ahead. *J Proteome* 169:41–57
 13. Ghatak A, Chaturvedi P, Weckwerth W (2017) Cereal crop proteomics: systemic analysis of crop drought stress responses towards marker-assisted selection breeding. *Front Plant Sci* 8: 757
 14. Glinski M, Weckwerth W (2006) The role of mass spectrometry in plant systems biology. *Mass Spectrom Rev* 25:173–214
 15. Paul P, Chaturvedi P, Selymes M, Ghatak A, Mesihovic A, Scharf KD, Weckwerth W, Simm S, Schleiff E (2016) The membrane proteome of male gametophyte in *Solanum lycopersicum*. *J Proteome* 131:48–60
 16. Weckwerth W (2003) Metabolomics in systems biology. *Annu Rev Plant Biol* 54:669–689

17. Weckwerth W (2008) Integration of metabolomics and proteomics in molecular plant physiology—coping with the complexity by data-dimensionality reduction. *Physiol Plant* 132: 176–189
18. Weckwerth W (2011) Green systems biology – from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *J Proteome* 75:284–305
19. Holtorf H, Guitton MC, Reski R (2002) Plant functional genomics. *Naturwissenschaften* 89:235–249
20. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2:343–372
21. Kitano H (2000) Perspectives on systems biology. *N Gener Comput* 18:199–216
22. Oliver DJ, Nikolau B, Wurtele ES (2002) Functional genomics: high-throughput mRNA, protein, and metabolite analyses. *Metab Eng* 4:98–106
23. Somerville C, Somerville S (1999) Plant functional genomics. *Science* 285:380–383
24. Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol Biol* 48:155–171
25. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. *Nat Biotechnol* 18:1157–1161
26. Weckwerth W, Fiehn O (2002) Can we discover novel pathways using metabolomic analysis? *Curr Opin Biotechnol* 13:156–160
27. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 19:1720–1730
28. Weckwerth W, Wenzel K, Fiehn O (2004) Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 4:78–83
29. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol* 22:245–252
30. Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol* 169: 453–468
31. Devaux PG, Horning MG, Hill RM, Horning EC (1971) O-benzoyloximes: derivatives for the study of ketosteroids by gas chromatography. Application to urinary steroids of the newborn human. *Anal Biochem* 41:70–82
32. Horning EC, Horning MG (1970) Metabolic profiles: chromatographic methods for isolation and characterization of a variety of metabolites in man. *Methods Med Res* 12:369–371
33. Horning EC, Horning MG (1971) Metabolic profiles: gas-phase methods for analysis of metabolites. *Clin Chem* 17:802–809
34. Cunnick WR, Cromie JB, Cortell R, Wright B, Beach E, Seltzer F, Miller S (1972) Value of biochemical profiling in a periodic health examination program: analysis of 1,000 cases. *Bull N Y Acad Med* 48:5–22
35. Mroczek WJ (1972) Biochemical profiling and the natural history of hypertensive diseases. *Circulation* 45:1332–1333
36. Vrbanc JJ, Braselton Jr WE, Holland JF, Sweeley CC (1982) Automated qualitative and quantitative metabolic profiling analysis of urinary steroids by a gas chromatography-mass spectrometry-data system. *J Chromatogr* 239:265–276
37. Niwa T (1986) Metabolic profiling with gas chromatography-mass spectrometry and its application to clinical medicine. *J Chromatogr* 379:313–345
38. Bales JR, Bell JD, Nicholson JK, Sadler PJ, Timbrell JA, Hughes RD, Bennett PN, Williams R (1988) Metabolic profiling of body fluids by proton NMR: self-poisoning episodes with paracetamol (acetaminophen). *Magn Reson Med* 6:300–306
39. Bales JR, Higham DP, Howe I, Nicholson JK, Sadler PJ (1984) Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. *Clin Chem* 30:426–432
40. Sauter H, Lauer M, Fritsch H (1991) Metabolic profiling of plants – a new diagnostic-technique. *ACS Symp Ser* 443:288–299
41. Oliver SG, Winson MK, Kell DB, Baganz F (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol* 16:373–378

42. Trethewey RN, Krotzky AJ, Willmitzer L (1999) Metabolic profiling: a Rosetta stone for genomics? *Curr Opin Plant Biol* 2:83–85
43. Tweeddale H, Noltley-Mcrobbs L, Ferenci T (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool (“metabolome”) analysis. *J Bacteriol* 180:5109–5116
44. Sumner LW, Duran AL, Huhman DV, Smith JT (2002) Metabolomics: a developing and integral component in functional genomic studies of *Medicago truncatula*. *Phytochem Genom Post-Genom Era* 36(31–61):258
45. Allwood JW, Goodacre R (2010) An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochem Anal* 21:33–47
46. Moco S, Bino RJ, De Vos RCH, Vervoort J (2007) Metabolomics technologies and metabolite identification. *TrAC-Trends Analyt Chem* 26:855–866
47. Kikuchi J, Hirayama T (2007) Practical aspects of uniform stable isotope labeling of higher plants for heteronuclear NMR-based metabolomics. *Methods Mol Biol* 358:273–286
48. Kim HK, Choi YH, Verpoorte R (2010) NMR-based metabolomic analysis of plants. *Nat Protoc* 5:536–549
49. Kim HK, Choi YH, Verpoorte R (2011) NMR-based plant metabolomics: where do we stand, where do we go? *Trends Biotechnol* 29:267–275
50. Weckwerth W (2010) Metabolomics: an integral technique in systems biology. *Bioanalysis* 2:829–836
51. Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 62:817–836
52. Dunn WB, Ellis DI (2005) Metabolomics: current analytical platforms and methodologies. *TrAC-Trends Analyt Chem* 24:285–294
53. Tugizimana F, Piater L, Dubery I (2013) Plant metabolomics: a new frontier in phytochemical analysis. *S Afr J Sci* 109:11
54. Bligny R, Douce R (2001) NMR and plant metabolism. *Curr Opin Plant Biol* 4:191–196
55. Ratcliffe RG, Roscher A, Shachar-Hill Y (2001) Plant NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* 39:267–300
56. Griffin JL, Williams HJ, Sang E, Clarke K, Rae C, Nicholson JK (2001) Metabolic profiling of genetic disorders: a multitissue H-1 nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue. *Anal Biochem* 293:16–21
57. Brindle JT, Nicholson JK, Schofield PM, Grainger DJ, Holmes E (2003) Application of Chemometrics to H-1 NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst* 128:32–36
58. Solanky KS, Bailey NJC, Holmes E, Lindon JC, Davis AL, Mulder TPJ, Van Duynhoven JPM, Nicholson JK (2003) NMR-based metabolomic studies on the biochemical effects of epicatechin in the rat. *J Agric Food Chem* 51:4139–4145
59. Vaidyanathan S, Kell DB, Goodacre R (2002) Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. *J Am Soc Mass Spectrom* 13:118–128
60. Aharoni A, Ric De Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R, Goodenow DB (2002) Nontargeted metabolome analysis by use of fourier transform ion cyclotron mass spectrometry. *OMICS* 6:217–234
61. Gu HW, Huang YA, Carr PW (2011) Peak capacity optimization in comprehensive two dimensional liquid chromatography: a practical approach. *J Chromatogr A* 1218:64–73
62. Guiochon G, Marchetti N, Mriziq K, Shalliker RA (2008) Implementations of two-dimensional liquid chromatography. *J Chromatogr A* 1189:109–168
63. Jandera P (2012) Programmed elution in comprehensive two-dimensional liquid chromatography. *J Chromatogr A* 1255:112–129
64. Kempa S, Hummel J, Schwemmer T, Pietzke M, Strehmel N, Wienkoop S, Kopka J, Weckwerth W (2009) An automated GCxGC-TOF-MS protocol for batch-wise extraction and alignment of mass isotopomer matrixes from differential ¹³C-labelling experiments: a

- case study for photoautotrophic-mixotrophic grown *Chlamydomonas reinhardtii* cells. *J Basic Microbiol* 49:82–91
65. Ralston-Hooper K, Hopf A, Oh C, Zhang X, Adamec J, Sepulveda MS (2008) Development of GCxGC/TOF-MS metabolomics for use in ecotoxicological studies with invertebrates. *Aquat Toxicol* 88:48–52
 66. Morgenthal K, Wienkoop S, Scholz M, Selbig J, Weckwerth W (2005) Correlative GC-TOF-MS-based metabolite profiling and LC-MS-based protein profiling reveal time-related systemic regulation of metabolite-protein networks and improve pattern recognition for multiple biomarker selection. *Metabolomics* 1:109–121
 67. Morgenthal K, Wienkoop S, Wolschin F, Weckwerth W (2007) Integrative profiling of metabolites and proteins: improving pattern recognition and biomarker selection for systems level approaches. *Methods Mol Biol* 358:57–75
 68. Weckwerth W, Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today* 10:1551–1558
 69. Weckwerth W, Tolstikov V, Fiehn O (2001) Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS analysis reveals silent metabolic phenotypes. In: *Proceedings of the 49th ASMS conference on mass spectrometry and allied topics: American Society Of Mass Spectrometry Chicago*, pp 1–2
 70. Dunn WB (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Phys Biol* 5:011001
 71. Scherling C, Roscher C, Giavalisco P, Schulze ED, Weckwerth W (2010) Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. *PLoS One* 5:e12569
 72. Weckwerth W, Loureiro ME, Wenzel K, Fiehn O (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc Natl Acad Sci U S A* 101:7809–7814
 73. Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L (2004) Innovation – metabolite profiling: from diagnostics to systems biology. *Nat Rev Mol Cell Biol* 5:763–769
 74. Halket JM, Waterman D, Przyborowska AM, Patel RKP, Fraser PD, Bramley PM (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J Exp Bot* 56:219–243
 75. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat Protoc* 1:387–396
 76. Farre EM, Tiessen A, Roessner U, Geigenberger P, Trethewey RN, Willmitzer L (2001) Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids, and sugar alcohols in potato tubers using a nonaqueous fractionation method. *Plant Physiol* 127:685–700
 77. Roessner U, Willmitzer L, Fernie AR (2001) High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol* 127:749–764
 78. Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal Chem* 72:3573–3580
 79. Desbrosses GG, Kopka J, Udvardi MK (2005) Lotus japonicus metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. *Plant Physiol* 137:1302–1318
 80. Roessner-Tunali U, Hegemann B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR (2003) Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiol* 133:84–99
 81. Doerfler H, Lyon D, Nagele T, Sun X, Fagner L, Hadacek F, Egelhofer V, Weckwerth W (2013) Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics* 9:564–574

82. Doerfler H, Sun X, Wang L, Engelmeier D, Lyon D, Weckwerth W (2014) Mzgroup analyzer--predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. *PLoS One* 9:e96188
83. Meijon M, Feito I, Oravec M, Delatorre C, Weckwerth W, Majada J, Villedor L (2016) Exploring natural variation of *Pinus pinaster* Aiton using metabolomics: is it possible to identify the region of origin of a pine from its metabolites? *Mol Ecol* 25:959–976
84. Wang L, Nagele T, Doerfler H, Fragner L, Chaturvedi P, Nukarinen E, Bellaire A, Huber W, Weiszmann J, Engelmeier D, Ramsak Z, Gruden K, Weckwerth W (2016) System level analysis of cacao seed ripening reveals a sequential interplay of primary and secondary metabolism leading to polyphenol accumulation and preparation of stress resistance. *Plant J* 87:318–332
85. Wang L, Sun X, Weiszmann J, Weckwerth W (2017) System-level and granger network analysis of integrated proteomic and metabolomic dynamics identifies key points of grape berry development at the interface of primary and secondary metabolism. *Front Plant Sci* 8:1066
86. Tolstikov VV, Fiehn O (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal Biochem* 301:298–307
87. Tolstikov VV, Fiehn O, Tanaka N (2007) Application of liquid chromatography-mass spectrometry analysis in metabolomics: reversed-phase monolithic capillary chromatography and hydrophilic chromatography coupled to electrospray ionization-mass spectrometry. *Methods Mol Biol* 358:141–155
88. Ramautar R, Mayboroda OA, Somsen GW, De Jong GJ (2011) CE-MS for metabolomics: developments and applications in the period 2008–2010. *Electrophoresis* 32:52–65
89. Ramautar R, Somsen GW, De Jong GJ (2009) CE-MS in metabolomics. *Electrophoresis* 30:276–291
90. Soga T (2007) Capillary electrophoresis-mass spectrometry for metabolomics. *Methods Mol Biol* 358:129–137
91. Monton MR, Soga T (2007) Metabolome analysis by capillary electrophoresis-mass spectrometry. *J Chromatogr A* 1168:237–246. Discussion 236
92. Watanabe CK, Sato S, Yanagisawa S, Uesono Y, Terashima I, Noguchi K (2014) Effects of elevated CO₂ on levels of primary metabolites and transcripts of genes encoding respiratory enzymes and their diurnal patterns in *Arabidopsis thaliana*: possible relationships with respiratory rates. *Plant Cell Physiol* 55:341–357
93. Maruyama K, Urano K, Yoshiwara K, Morishita Y, Sakurai N, Suzuki H, Kojima M, Sakakibara H, Shibata D, Saito K, Shinozaki K, Yamaguchi-Shinozaki K (2014) Integrated analysis of the effects of cold and dehydration on rice metabolites, phytohormones, and gene transcripts. *Plant Physiol* 164:1759–1771
94. Hoehenwarter W, Larhlmi A, Hummel J, Egelhofer V, Selbig J, Van Dongen JT, Wienkoop S, Weckwerth W (2011) Mapa distinguishes genotype-specific variability of highly similar regulatory protein isoforms in potato tuber. *J Proteome Res* 10:2979–2991
95. Hoehenwarter W, Van Dongen JT, Wienkoop S, Steinfath M, Hummel J, Erban A, Sulpice R, Regierer B, Kopka J, Geigenberger P, Weckwerth W (2008) A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics* 8:4214–4225
96. 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* 8:1–13
97. Sun X, Weckwerth W (2013) Using COVAIn to analyze metabolomics data. In: Weckwerth W, Kahl G (eds) *The handbook of plant metabolomics*. Wiley-Blackwell, Weinheim, pp 305–320. <https://doi.org/10.1002/9783527669882.ch17>
98. Weckwerth W (2011) Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. *Anal Bioanal Chem* 400:1967–1978

99. Boccard J, Grata E, Thiocone A, Gauvrit JY, Lanteri P, Carrupt PA, Wolfender JL, Rudaz S (2007) Multivariate data analysis of rapid LC-TOF/MS experiments from *Arabidopsis thaliana* stressed by wounding. *Chemom Intell Lab Syst* 86:189–197
100. Liland KH (2011) Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis. *TrAC-Trends Anal Chem* 30:827–841
101. Jansen JJ, Smit S, Hoefsloot HCJ, Smilde AK (2010) The photographer and the greenhouse: how to analyse plant metabolomics data. *Phytochem Anal* 21:48–60
102. Van Den Berg RA, Rubingh CM, Westerhuis JA, Van Der Werf MJ, Smilde AK (2009) Metabolomics data exploration guided by prior knowledge. *Anal Chim Acta* 651:173–181
103. Vichi M, Saporta G (2009) Clustering and disjoint principal component analysis. *Comput Stat Data Anal* 53:3194–3208
104. Daub CO, Kloska S, Selbig J (2003) Metagenealyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics* 19:2332–2333
105. Xia JG, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS (2012) Metaboanalyst 2.0-A comprehensive server for metabolomic data analysis. *Nucleic Acids Res* 40:W127–W133
106. Barupal DK, Haldiya PK, Wohlgemuth G, Kind T, Kothari SL, Pinkerton KE, Fiehn O (2012) MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics* 13:99
107. Kastenmuller G, Romisch-Margl W, Wägele B, Altmaier E, Suhre K (2011) metaP-server: a web-based metabolomics data analysis tool. *J Biomed Biotechnol* 2011:7
108. Neuweger H, Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K, Stoye J, Goesmann A (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics* 24:2726–2732
109. Rojas-Cherto M, Van Vliet M, Peironcely JE, Van Doorn R, Kooyman M, Te Beek T, Van Driel MA, Hankemeier T, Reijmers T (2012) MetiTree: a web application to organize and process high-resolution multi-stage mass spectrometry metabolomics data. *Bioinformatics* 28:2707–2709
110. Nagele T, Mair A, Sun X, Fragner L, Teige M, Weckwerth W (2014) Solving the differential biochemical Jacobian from metabolomics covariance data. *PLoS One* 9:e92299
111. Nagele T, Weckwerth W (2013) A workflow for mathematical modeling of subcellular metabolic pathways in leaf metabolism of *Arabidopsis thaliana*. *Front Plant Sci* 4:541
112. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40:D109–D114
113. Zhang PF, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol* 138:27–37
114. Schreiber F, Colmsee C, Czauderna T, Grafahrend-Belau E, Hartmann A, Junker A, Junker BH, Klapperstuck M, Scholz U, Weise S (2012) MetaCrop 2.0: managing and exploring information about crop plant metabolism. *Nucleic Acids Res* 40:D1173–D1177
115. Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, Bridge A, Bougueleret L, Xenarios I, Viari A (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res* 40:D761–D769
116. Frolkis A, Knox C, Lim E, Jewison T, Law V, Hau DD, Liu P, Gautam B, Ly S, Guo AC, Xia JG, Liang YJ, Shrivastava S, Wishart DS (2010) SMPDB: the small molecule pathway database. *Nucleic Acids Res* 38:D480–D487
117. Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ* 32:1211–1229
118. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-Mcintyre S, Begley P, Carroll K, Broadhurst D, Tseng A, Swainston N, Spasic I, Goodacre R, Kell DB (2009) Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst* 134:1322–1332
119. Okazaki Y, Saito K (2012) Recent advances of metabolomics in plant biotechnology. *Plant Biotechnol Rep* 6:1–15

120. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3:211–221
121. Hoehenwarter W, Chen Y, Recuenco-Munoz L, Wienkoop S, Weckwerth W (2011) Functional analysis of proteins and protein species using shotgun proteomics and linear mathematics. *Amino Acids* 41:329–341
122. Klose J, Kobalz U (1995) 2-dimensional electrophoresis of proteins – an updated protocol and implications for a functional-analysis of the genome. *Electrophoresis* 16:1034–1059
123. Morgenthal K, Weckwerth W, Steuer R (2006) Metabolomic networks in plants: transitions from pattern recognition to biological interpretation. *Biosystems* 83:108–117
124. Smith RD, Loo JA, Loo RRO, Busman M, Udseth HR (1991) Principles and practice of electrospray ionization – mass-spectrometry for large polypeptides and proteins. *Mass Spectrom Rev* 10:359–451
125. Weckwerth W, Kahl G (2013) *The handbook of plant metabolomics*. Wiley, Hoboken
126. Gorinstein S, Zenser M, Vargasalbores F, Ochoa JL (1995) Classification of 7 species of Cactaceae based on their chemical and biochemical-properties. *Biosci Biotechnol Biochem* 59:2022–2027
127. Bednarek P, Franski R, Kerhoas L, Einhorn J, Wojtaszek P, Stobiecki M (2001) Profiling changes in metabolism of isoflavonoids and their conjugates in *Lupinus albus* treated with biotic elicitor. *Phytochemistry* 56:77–85
128. Lois R (1994) Accumulation of Uv-absorbing flavonoids induced by Uv-B radiation in *Arabidopsis-thaliana* L.1. Mechanisms of Uv-resistance in *Arabidopsis*. *Planta* 194:498–503
129. Harrison MJ, Dixon RA (1993) Isoflavonoid accumulation and expression of defense gene transcripts during the establishment of vesicular-Arbuscular mycorrhizal associations in roots of *Medicago-truncatula*. *Mol Plant-Microbe Interact* 6:643–654
130. Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, Willmitzer L, Zamir D, Fernie AR (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat Biotechnol* 24:447–454
131. Nägele T, Fragner L, Chaturvedi P, Ghatak A, Weckwerth W (2017) Pollen metabolome dynamics: biochemistry, regulation and analysis. Springer, Cham
132. Roldan MVG, Engel B, De Vos RCH, Vereijken P, Astola L, Groenenboom M, Van De Geest H, Bovy A, Molenaar J, Van Eeuwijk F, Hall RD (2014) Metabolomics reveals organ-specific metabolic rearrangements during early tomato seedling development. *Metabolomics* 10:958–974
133. Dietrich CR, Han G, Chen M, Berg RH, Dunn TM, Cahoon EB (2008) Loss-of-function mutations and inducible RNAi suppression of *Arabidopsis* LCB2 genes reveal the critical role of sphingolipids in gametophytic and sporophytic cell viability. *Plant J* 54:284–298
134. Allwood JW, Ellis DI, Goodacre R (2008) Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period. *Physiol Plant* 132:117–135
135. Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci* 9:418–425
136. Kim HK, Verpoorte R (2010) Sample preparation for plant metabolomics. *Phytochem Anal* 21:4–13
137. Allwood JW, Ellis DI, Goodacre R (2008) Metabolomic technologies and their application to the study of plants and plant-host interactions. *Physiol Plant* 132:117–135
138. Valledor L, Escandon M, Meijon M, Nukarinen E, Canal MJ, Weckwerth W (2014) A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms. *Plant J* 79:173–180
139. Wienkoop S, Morgenthal K, Wolschin F, Scholz M, Selbig J, Weckwerth W (2008) Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in *Arabidopsis thaliana*. *Mol Cell Proteomics* 7:1725–1736

140. Antonio C, Pinheiro C, Chaves MM, Ricardo CP, Ortuno MF, Thomas-Oates J (2008) Analysis of carbohydrates in *Lupinus albus* stems on imposition of water deficit, using porous graphitic carbon liquid chromatography-electrospray ionization mass spectrometry (Vol 1187, Pg 111, 2008). *J Chromatogr A* 1201:132–132
141. Gechev TS, Benina M, Obata T, Tohge T, Sujeeth N, Minkov I, Hille J, Temanni MR, Marriott AS, Bergstrom E, Thomas-Oates J, Antonio C, Mueller-Roeber B, Schippers JH, Fernie AR, Toneva V (2013) Molecular mechanisms of desiccation tolerance in the resurrection glacial relic *Haberlea rhodopensis*. *Cell Mol Life Sci* 70:689–709
142. Erxleben A, Gessler A, Vervliet-Scheebaum M, Reski R (2012) Metabolite profiling of the moss *Physcomitrella patens* reveals evolutionary conservation of osmoprotective substances. *Plant Cell Rep* 31:427–436
143. Bowne JB, Erwin TA, Juttner J, Schnurbusch T, Langridge P, Bacic A, Roessner U (2012) Drought responses of leaf tissues from wheat cultivars of differing drought tolerance at the metabolite level. *Mol Plant* 5:418–429
144. Do PT, Degenkolbe T, Erban A, Heyer AG, Kopka J, Kohl KI, Hinch DK, Zuther E (2013) Dissecting rice polyamine metabolism under controlled long-term drought stress. *PLoS One* 8: e60325
145. Yang JC, Zhang JH, Liu K, Wang ZQ, Liu LJ (2007) Involvement of polyamines in the drought resistance of rice. *J Exp Bot* 58:1545–1555
146. Urano K, Maruyama K, Ogata Y, Morishita Y, Takeda M, Sakurai N, Suzuki H, Saito K, Shibata D, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2009) Characterization of the ABA-regulated global responses to dehydration in *Arabidopsis* by metabolomics. *Plant J* 57:1065–1078
147. Skirycz A, De Bodt S, Obata T, De Clercq I, Claeys H, De Rycke R, Andriankaja M, Van Aken O, Van Breusegem F, Fernie AR, Inze D (2010) Developmental stage specificity and the role of mitochondrial metabolism in the response of *Arabidopsis* leaves to prolonged mild osmotic stress. *Plant Physiol* 152:226–244
148. Semel Y, Schauer N, Roessner U, Zamir D, Fernie AR (2007) Metabolite analysis for the comparison of irrigated and non-irrigated field grown tomato of varying genotype. *Metabolomics* 3:289–295
149. Witt S, Galicia L, Lisec J, Cairns J, Tiessen A, Araus JL, Palacios-Rojas N, Fernie AR (2012) Metabolic and phenotypic responses of greenhouse-grown maize hybrids to experimentally controlled drought stress. *Mol Plant* 5:401–417
150. Sicher RC, Barnaby JY (2012) Impact of carbon dioxide enrichment on the responses of maize leaf transcripts and metabolites to water stress. *Physiol Plant* 144:238–253
151. Silvente S, Sobolev AP, Lara M (2012) Metabolite adjustments in drought tolerant and sensitive soybean genotypes in response to water stress. *PLoS One* 7:e38554
152. Cook D, Fowler S, Fiehn O, Thomashow MF (2004) A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of *Arabidopsis*. *Proc Natl Acad Sci U S A* 101:15243–15248
153. Espinoza C, Degenkolbe T, Caldana C, Zuther E, Leisse A, Willmitzer L, Hinch DK, Hannah MA (2010) Interaction with diurnal and circadian regulation results in dynamic metabolic and transcriptional changes during cold acclimation in *Arabidopsis*. *PLoS One* 5:e14101
154. Kaplan F, Kopka J, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of *Arabidopsis*. *Plant Physiol* 136:4159–4168
155. Kaplan F, Kopka J, Sung DY, Zhao W, Popp M, Porat R, Guy CL (2007) Transcript and metabolite profiling during cold acclimation of *Arabidopsis* reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. *Plant J* 50:967–981
156. Korn M, Gartner T, Erban A, Kopka J, Selbig J, Hinch DK (2010) Predicting *Arabidopsis* freezing tolerance and heterosis in freezing tolerance from metabolite composition. *Mol Plant* 3:224–235
157. Nagler M, Nukarinen E, Weckwerth W, Nagele T (2015) Integrative molecular profiling indicates a central role of transitory starch breakdown in establishing a stable C/N homeostasis during cold acclimation in two natural accessions of *Arabidopsis thaliana*. *BMC Plant Biol* 15:284

158. Paupiere MJ, Muller F, Li HJ, Rieu I, Tikunov YM, Visser RGF, Bovy AG (2017) Untargeted metabolomic analysis of tomato pollen development and heat stress response. *Plant Reprod* 30:81–94
159. Qi XL, Xu WG, Zhang JZ, Guo R, Zhao MZ, Hu L, Wang HW, Dong HB, Li Y (2017) Physiological characteristics and metabolomics of transgenic wheat containing the maize C-4 phosphoenolpyruvate carboxylase (PEPC) gene under high temperature stress. *Protoplasma* 254:1017–1030
160. Sun CX, Gao XX, Li MQ, Fu JQ, Zhang YL (2016) Plastic responses in the metabolome and functional traits of maize plants to temperature variations. *Plant Biol* 18:249–261
161. Brosche M, Vinocur B, Alatalo ER, Lamminmaki A, Teichmann T, Ottow EA, Djilianov D, Afif D, Bogeat-Triboulet MB, Altman A, Polle A, Dreyer E, Rudd S, Lars P, Auvinen P, Kangasjarvi J (2005) Gene expression and metabolite profiling of *Populus euphratica* growing in the Negev desert. *Genome Biol* 6:R101
162. Cramer GR, Ergul A, Grimplet J, Tillett RL, Tattersall EAR, Bohlman MC, Vincent D, Sonderegger J, Evans J, Osborne C, Quilici D, Schlauch KA, Schooley DA, Cushman JC (2007) Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Funct Integr Genomics* 7:111–134
163. Gagneul D, Ainouche A, Duhaze C, Lukan R, Larher FR, Bouchereau A (2007) A reassessment of the function of the so-called compatible solutes in the halophytic Plumbaginaceae *Limonium latifolium*. *Plant Physiol* 144:1598–1611
164. Gavaghan CL, Li JV, Hadfield ST, Hole S, Nicholson JK, Wilson ID, Howe PWA, Stanley PD, Holmes E (2011) Application of NMR-based metabolomics to the investigation of salt stress in maize (*Zea mays*). *Phytochem Anal* 22:214–224
165. Gong QQ, Li PH, Ma SS, Rupassara SI, Bohnert HJ (2005) Salinity stress adaptation competence in the extremophile *Thellungiella halophila* in comparison with its relative *Arabidopsis thaliana*. *Plant J* 44:826–839
166. Kim JK, Bamba T, Harada K, Fukusaki E, Kobayashi A (2007) Time-course metabolic profiling in *Arabidopsis thaliana* cell cultures after salt stress treatment. *J Exp Bot* 58:415–424
167. Gupta P, De B (2017) Metabolomics analysis of rice responses to salinity stress revealed elevation of serotonin, and gentisic acid levels in leaves of tolerant varieties. *Plant Signal Behav* 12(7):e1335845
168. Liu D, Ford KL, Roessner U, Natera S, Cassin AM, Patterson JH, Bacic A (2013) Rice suspension cultured cells are evaluated as a model system to study salt responsive networks in plants using a combined proteomic and metabolomic profiling approach. *Proteomics* 13:2046–2062
169. Sanchez DH, Lippold F, Redestig H, Hannah MA, Erban A, Kramer U, Kopka J, Udvardi MK (2008) Integrative functional genomics of salt acclimatization in the model legume *Lotus japonicus*. *Plant J* 53:973–987
170. Sanchez DH, Pieckenstain FL, Escaray F, Erban A, Kraemer U, Udvardi MK, Kopka J (2011) Comparative ionomics and metabolomics in extremophile and glycophytic *Lotus* species under salt stress challenge the metabolic pre-adaptation hypothesis. *Plant Cell Environ* 34: 605–617
171. Widodo, Patterson JH, Newbigin E, Tester M, Bacic A, Roessner U (2009) Metabolic responses to salt stress of barley (*Hordeum vulgare* L.) cultivars, Sahara and Clipper, which differ in salinity tolerance. *J Exp Bot* 60:4089–4103
172. Wu D, Cai S, Chen M, Ye L, Chen Z, Zhang H, Dai F, Wu F, Zhang G (2013) Tissue metabolic responses to salt stress in wild and cultivated barley. *PLoS One* 8:e55431
173. Baxter CJ, Redestig H, Schauer N, Repsilber D, Patil KR, Nielsen J, Selbig J, Liu JL, Fernie AR, Sweetlove LJ (2007) The metabolic response of heterotrophic *Arabidopsis* cells to oxidative stress. *Plant Physiol* 143:312–325
174. Lehmann M, Schwarzlander M, Obata T, Sirikantaramas S, Burow M, Olsen CE, Tohge T, Fricker MD, Moller BL, Fernie AR, Sweetlove LJ, Laxa M (2009) The metabolic response of *Arabidopsis* roots to oxidative stress is distinct from that of heterotrophic cells in culture and highlights a complex relationship between the levels of transcripts, metabolites, and flux. *Mol Plant* 2:390–406

175. Ishikawa T, Takahara K, Hirabayashi T, Matsumura H, Fujisawa S, Terauchi R, Uchimiya H, Kawai-Yamada M (2010) Metabolome analysis of response to oxidative stress in Rice suspension cells overexpressing cell death suppressor Bax inhibitor-1. *Plant Cell Physiol* 51:9–20
176. Komatsu S, Yamamoto A, Nakamura T, Nakamura T, Nouri MZ, Nanjo Y, Nishizawa K, Furukawa K (2011) Comprehensive analysis of mitochondria in roots and hypocotyls of soybean under flooding stress using proteomics and metabolomics techniques. *J Proteome Res* 10:3993–4004
177. Komatsu S, Nakamura T, Sugimoto Y, Sakamoto K (2014) Proteomic and metabolomic analyses of soybean root tips under flooding stress. *Protein Pept Lett* 21:865–884
178. Kusano M, Tabuchi M, Fukushima A, Funayama K, Diaz C, Kobayashi M, Hayashi N, Tsuchiya YN, Takahashi H, Kamata A, Yamaya T, Saito K (2011) Metabolomics data reveal a crucial role of cytosolic glutamine Synthetase 1;1 in coordinating metabolic balance in rice. *Plant J* 66:456–466
179. Urbanczyk-Wochniak E, Fernie AR (2005) Metabolic profiling reveals altered nitrogen nutrient regimes have diverse effects on the metabolism of hydroponically-grown tomato (*Solanum lycopersicum*) plants. *J Exp Bot* 56:309–321
180. Tschöp H, Gibon Y, Carillo P, Armengaud P, Szczowka M, Nunes-Nesi A, Fernie AR, Koehl K, Stitt M (2009) Adjustment of growth and central metabolism to a mild but sustained nitrogen-limitation in *Arabidopsis*. *Plant Cell Environ* 32:300–318
181. Nikiforova VJ, Daub CO, Hesse H, Willmitzer L, Hoefgen R (2005) Integrative metabolite network with implemented causality deciphers informational fluxes of Sulphur stress response. *J Exp Bot* 56:1887–1896
182. Morcuende R, Bari R, Gibon Y, Zheng WM, Pant BD, Blasing O, Usadel B, Czechowski T, Udvardi MK, Stitt M, Scheible WR (2007) Genome-wide reprogramming of metabolism and regulatory networks of *Arabidopsis* in response to phosphorus. *Plant Cell Environ* 30:85–112
183. Hernandez G, Valdes-Lopez O, Ramirez M, Goffard N, Weiller G, Aparicio-Fabre R, Fuentes SI, Erban A, Kopka J, Udvardi MK, Vance CP (2009) Global changes in the transcript and metabolic profiles during symbiotic nitrogen fixation in phosphorus-stressed common bean plants. *Plant Physiol* 151:1221–1238
184. Hernandez G, Ramirez M, Valdes-Lopez O, Tesfaye M, Graham MA, Czechowski T, Schlereth A, Wandrey M, Erban A, Cheung F, Wu HC, Lara M, Town CD, Kopka J, Udvardi MK, Vance CP (2007) Phosphorus stress in common bean: root transcript and metabolic responses. *Plant Physiol* 144:752–767
185. Huang CY, Roessner U, Eickmeier I, Genc Y, Callahan DL, Shirley N, Langridge P, Bacic A (2008) Metabolite profiling reveals distinct changes in carbon and nitrogen metabolism in phosphate-deficient barley plants (*Hordeum vulgare* L.) *Plant Cell Physiol* 49:691–703
186. Armengaud P, Sulpice R, Miller AJ, Stitt M, Amtmann A, Gibon Y (2009) Multilevel analysis of primary metabolism provides new insights into the role of potassium nutrition for glycolysis and nitrogen assimilation in *Arabidopsis* roots. *Plant Physiol* 150:772–785
187. Jahangir M, Abdel-Farid IB, Choi YH, Verpoorte R (2008) Metal ion-inducing metabolite accumulation in *Brassica rapa*. *J Plant Physiol* 165:1429–1437
188. Sun XM, Zhang JX, Zhang HJ, Ni YW, Zhang Q, Chen JP, Guan YF (2010) The responses of *Arabidopsis thaliana* to cadmium exposure explored via metabolite profiling. *Chemosphere* 78:840–845
189. Agarrwal R, Bentur JS, Nair S (2014) Gas chromatography mass spectrometry based metabolic profiling reveals biomarkers involved in rice-gall midge interactions. *J Integr Plant Biol* 56:837–848
190. Sana TR, Fischer S, Wohlgemuth G, Katrekar A, Jung KH, Ronald PC, Fiehn O (2010) Metabolomic and transcriptomic analysis of the rice response to the bacterial blight pathogen *Xanthomonas oryzae* pv. *oryzae*. *Metabolomics* 6:451–465
191. Parker D, Beckmann M, Zubair H, Enot DP, Caracuel-Rios Z, Overy DP, Snowdon S, Talbot NJ, Draper J (2009) Metabolomic analysis reveals a common pattern of metabolic re-programming during invasion of three host plant species by *Magnaporthe grisea*. *Plant J* 59:723–737

192. Gunnaiah R, Kushalappa AC, Duggavathi R, Fox S, Somers DJ (2012) Integrated metabolo-proteomic approach to decipher the mechanisms by which wheat QTL (Fhb1) contributes to resistance against *Fusarium graminearum*. *PLoS One* 7:e40695
193. Chaves MM, Maroco JP, Pereira JS (2003) Understanding plant responses to drought – from genes to the whole plant. *Funct Plant Biol* 30:239–264
194. Chaves MM, Oliveira MM (2004) Mechanisms underlying plant resilience to water deficits: prospects for water-saving agriculture. *J Exp Bot* 55:2365–2384
195. Hare PD, Cress WA, Van Staden J (1998) Dissecting the roles of osmolyte accumulation during stress. *Plant Cell Environ* 21:535–553
196. Yancey PH (2005) Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses. *J Exp Biol* 208:2819–2830
197. Bitrian M, Zarza X, Altabella T, Tiburcio AF, Alcazar R (2012) Polyamines under abiotic stress: metabolic crossroads and hormonal crosstalks in plants. *Metabolites* 2:516–528
198. Gill SS, Tuteja N (2010) Polyamines and abiotic stress tolerance in plants. *Plant Signal Behav* 5:26–33
199. Skirycz A, Inze D (2010) More from less: plant growth under limited water. *Curr Opin Biotechnol* 21:197–203
200. Guy CL (1990) Cold-acclimation and freezing stress tolerance – role of protein-metabolism. *Annu Rev Plant Physiol Plant Mol Biol* 41:187–223
201. Le Gall H, Fontaine JX, Molinie R, Pelloux J, Mesnard F, Gillet F, Fliniaux O (2017) NMR-based metabolomics to study the cold-acclimation strategy of two *Miscanthus* genotypes. *Phytochem Anal* 28:58–67
202. Hauser F, Horie T (2010) A conserved primary salt tolerance mechanism mediated by HKT transporters: a mechanism for sodium exclusion and maintenance of high K⁺/Na⁺ ratio in leaves during salinity stress. *Plant Cell Environ* 33:552–565
203. Munns R, Tester M (2008) Mechanisms of salinity tolerance. *Annu Rev Plant Biol* 65(59): 651–681
204. Sanchez DH, Siahpoosh MR, Roessner U, Udvardi M, Kopka J (2008) Plant metabolomics reveals conserved and divergent metabolic responses to salinity. *Physiol Plant* 132:209–219
205. Inan G, Zhang Q, Li PH, Wang ZL, Cao ZY, Zhang H, Zhang CQ, Quist TM, Goodwin SM, Zhu JH, Shi HH, Damsz B, Charbaji T, Gong QQ, Ma SS, Fredricksen M, Galbraith DW, Jenks MA, Rhodes D, Hasegawa PM, Bohnert HJ, Joly RJ, Bressan RA, Zhu JK (2004) Salt stress. A halophyte and cryophyte *Arabidopsis* relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant Physiol* 135:1718–1737
206. Johnson HE, Broadhurst D, Goodacre R, Smith AR (2003) Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry* 62:919–928
207. Shulaev V, Cortes D, Miller G, Mittler R (2008) Metabolomics for plant stress response. *Physiol Plant* 132:199–208
208. Mittler R (2002) Oxidative stress, antioxidants and stress tolerance. *Trends Plant Sci* 7: 405–410
209. Suzuki N, Koussevitzky S, Mittler R, Miller G (2012) ROS and redox signalling in the response of plants to abiotic stress. *Plant Cell Environ* 35:259–270
210. Morgan MJ, Lehmann M, Schwarzlander M, Baxter CJ, Sienkiewicz-Porzucek A, Williams TCR, Schauer N, Fernie AR, Fricker MD, Ratcliffe RG, Sweetlove LJ, Finkemeier I (2008) Decrease in manganese superoxide dismutase leads to reduced root growth and affects Tricarboxylic acid cycle flux and mitochondrial redox homeostasis. *Plant Physiol* 147: 101–114
211. Jackson MB, Ishizawa K, Ito O (2009) Evolution and mechanisms of plant tolerance to flooding stress. *Ann Bot* 103:137–142
212. Hoefgen R, Nikiforova VJ (2008) Metabolomics integrated with transcriptomics: assessing systems response to sulfur-deficiency stress. *Physiol Plant* 132:190–198

213. Nikiforova VJ, Kopka J, Tolstikov V, Fiehn O, Hopkins L, Hawkesford MJ, Hesse H, Hoefgen R (2005) Systems rebalancing of metabolism in response to Sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants. *Plant Physiol* 138:304–318
214. Sharma SS, Dietz KJ (2009) The relationship between metal toxicity and cellular redox imbalance. *Trends Plant Sci* 14:43–50
215. Balmerl D, Flors V, Glauser G, Mauch-Mani B (2013) Metabolomics of cereals under biotic stress: current knowledge and techniques. *Front Plant Sci* 4:82
216. Obata T, Fernie AR (2012) The use of metabolomics to dissect plant responses to abiotic stresses. *Cell Mol Life Sci* 69:3225–3243
217. Rizhsky L, Liang HJ, Shuman J, Shulaev V, Davletova S, Mittler R (2004) When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. *Plant Physiol* 134:1683–1696
218. Wulff-Zottele C, Gatzke N, Kopka J, Orellana A, Hoefgen R, Fisahn J, Hesse H (2010) Photosynthesis and metabolism interact during acclimation of Arabidopsis thaliana to high irradiance and sulphur depletion. *Plant Cell Environ* 33:1974–1988
219. Krasensky J, Jonak C (2012) Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J Exp Bot* 63:1593–1608
220. Empadinhas N, Da Costa MS (2008) Osmoadaptation mechanisms in prokaryotes: distribution of compatible solutes. *Int Microbiol* 11:151–161
221. Rathinasabapathi B (2000) Metabolic engineering for stress tolerance: installing osmo-protectant synthesis pathways. *Ann Bot* 86:709–716
222. Barnett NM, Naylor AW (1966) Amino acid and protein metabolism in Bermuda grass during water stress. *Plant Physiol* 41:1222
223. Verbruggen N, Hermans C (2008) Proline accumulation in plants: a review. *Amino Acids* 35:753–759
224. Verslues PE, Agarwal M, Katiyar-Agarwal S, Zhu J, Zhu JK (2006) Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. (Vol 45, Pg 523, 2006). *Plant J* 46:1092–1092
225. Verslues PE, Sharma S (2010) Proline metabolism and its implications for plant-environment interaction. *Arabidopsis Book* 8:e0140
226. Kishor PBK, Hong ZL, Miao GH, Hu CAA, Verma DPS (1995) Overexpression of delta-pyrroline-5-carboxylate synthetase increases proline production and confers osmotolerance in transgenic plants. *Plant Physiol* 108:1387–1394
227. Szekely G, Abraham E, Cselo A, Rigo G, Zsigmond L, Csiszar J, Ayaydin F, Strizhov N, Jasik J, Schmelzer E, Koncz C, Szabados L (2008) Duplicated P5cs genes of Arabidopsis play distinct roles in stress regulation and developmental control of proline biosynthesis. *Plant J* 53:11–28
228. Nanjo T, Kobayashi M, Yoshiba Y, Sanada Y, Wada K, Tsukaya H, Kakubari Y, Yamaguchi-Shinozaki K, Shinozaki K (1999) Biological functions of proline in morphogenesis and osmotolerance revealed in antisense transgenic Arabidopsis thaliana. *Plant J* 18:185–193
229. Roosens NH, Al Bitar F, Loenders K, Angenon G, Jacobs M (2002) Overexpression of ornithine-delta-aminotransferase increases proline biosynthesis and confers osmotolerance in transgenic plants. *Mol Breed* 9:73–80
230. Kaplan F, Guy CL (2004) Beta-amylase induction and the protective role of maltose during temperature shock. *Plant Physiol* 135:1674–1684
231. Kempa S, Krasensky J, Dal Santo S, Kopka J, Jonak C (2008) A central role of abscisic acid in stress-regulated carbohydrate metabolism. *PLoS One* 3:e3935
232. Renault H, Roussel V, El Amrani A, Arzel M, Renault D, Bouchereau A, Deleu C (2010) The Arabidopsis pop2-1 mutant reveals the involvement of GABA transaminase in salt stress tolerance. *BMC Plant Biol* 10:20
233. Fait A, Fromm H, Walter D, Galili G, Fernie AR (2008) Highway or byway: the metabolic role of the GABA shunt in plants. *Trends Plant Sci* 13:14–19
234. Liu CL, Zhao L, Yu GH (2011) The dominant glutamic acid metabolic flux to produce gamma-amino butyric acid over proline in Nicotiana tabacum leaves under water stress relates to its significant role in antioxidant activity. *J Integr Plant Biol* 53:608–618

235. Song HM, Xu XB, Wang H, Wang HZ, Tao YZ (2010) Exogenous gamma-aminobutyric acid alleviates oxidative damage caused by aluminium and proton stresses on barley seedlings. *J Sci Food Agric* 90:1410–1416
236. Bouche N, Fait A, Bouchez D, Moller SG, Fromm H (2003) Mitochondrial succinic-semialdehyde dehydrogenase of the gamma-aminobutyrate shunt is required to restrict levels of reactive oxygen intermediates in plants. *Proc Natl Acad Sci U S A* 100:6843–6848
237. Cromwell BT, Rennie SD (1953) The biosynthesis and metabolism of betaines in plants. I. The estimation and distribution of Glycinebetaine (betaine) in *Beta-vulgaris* L and other plants. *Biochem J* 55:189–192
238. Chen TH, Murata N (2011) Glycinebetaine protects plants against abiotic stress: mechanisms and biotechnological applications. *Plant Cell Environ* 34:1–20
239. Goel D, Singh AK, Yadav V, Babbar SB, Murata N, Bansal KC (2011) Transformation of tomato with a bacterial coda gene enhances tolerance to salt and water stresses. *J Plant Physiol* 168:1286–1294
240. Park EJ, Jeknic Z, Sakamoto A, Denoma J, Yuwansiri R, Murata N, Chen TH (2004) Genetic engineering of glycinebetaine synthesis in tomato protects seeds, plants, and flowers from chilling damage. *Plant J* 40:474–487
241. Waditee R, Bhuiyan MN, Rai V, Aoki K, Tanaka Y, Hibino T, Suzuki S, Takano J, Jagendorf AT, Takabe T, Takabe T (2005) Genes for direct methylation of glycine provide high levels of glycinebetaine and abiotic-stress tolerance in *Synechococcus* and *Arabidopsis*. *Proc Natl Acad Sci U S A* 102:1318–1323
242. Bianchi G, Gamba A, Limiroli R, Pozzi N, Elster R, Salamini F, Bartels D (1993) The unusual sugar composition in leaves of the resurrection plant *Myrothamnus-flabellifolia*. *Physiol Plant* 87:223–226
243. Paul MJ, Primavesi LF, Jhurrea D, Zhang YH (2008) Trehalose metabolism and signaling. *Annu Rev Plant Biol* 59:417–441
244. Guy C, Kaplan F, Kopka J, Selbig J, Hinch DK (2008) Metabolomics of temperature stress. *Physiol Plant* 132:220–235
245. Iordachescu M, Imai R (2008) Trehalose biosynthesis in response to abiotic stresses. *J Integr Plant Biol* 50:1223–1229
246. Li HW, Zang BS, Deng XW, Wang XP (2011) Overexpression of the trehalose-6-phosphate synthase gene *OsTPS1* enhances abiotic stress tolerance in rice. *Planta* 234:1007–1018
247. Suzuki N, Bajad S, Shuman J, Shulaev V, Mittler R (2008) The transcriptional co-activator MBF1c is a key regulator of thermotolerance in *Arabidopsis thaliana*. *J Biol Chem* 283:9269–9275
248. Panikulangara TJ, Eggers-Schumacher G, Wunderlich M, Stransky H, Schoffl F (2004) Galactinol synthase1. A novel heat shock factor target gene responsible for heat-induced synthesis of raffinose family oligosaccharides in *Arabidopsis*. *Plant Physiol* 136:3148–3158
249. Peterbauer T, Richter A (2001) Biochemistry and physiology of raffinose family oligosaccharides and galactosyl cyclitols in seeds. *Seed Sci Res* 11:185–197
250. Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J* 29:417–426
251. Nishizawa A, Yabuta Y, Yoshida E, Maruta T, Yoshimura K, Shigeoka S (2006) *Arabidopsis* heat shock transcription factor A2 as a key regulator in response to several types of environmental stress. *Plant J* 48:535–547

Epigenetics and Epigenomics of Plants

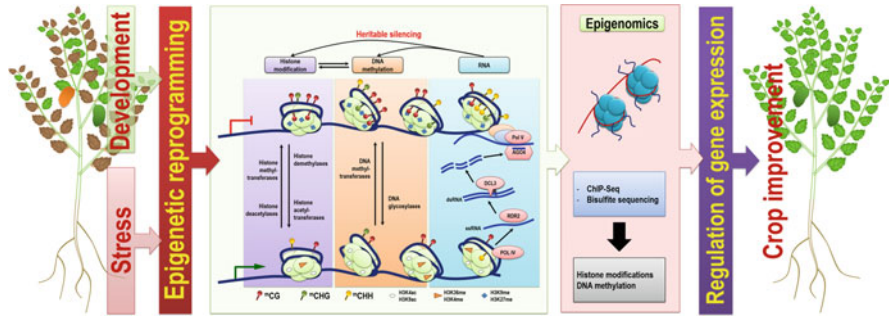


**Chandra Bhan Yadav, Garima Pandey, Mehanathan Muthamilarasan,
and Manoj Prasad**

Abstract The genetic material DNA in association with histone proteins forms the complex structure called chromatin, which is prone to undergo modification through certain epigenetic mechanisms including cytosine DNA methylation, histone modifications, and small RNA-mediated methylation. Alterations in chromatin structure lead to inaccessibility of genomic DNA to various regulatory proteins such as transcription factors, which eventually modulates gene expression. Advancements in high-throughput sequencing technologies have provided the opportunity to study the epigenetic mechanisms at genome-wide levels. Epigenomic studies using high-throughput technologies will widen the understanding of mechanisms as well as functions of regulatory pathways in plant genomes, which will further help in manipulating these pathways using genetic and biochemical approaches. This technology could be a potential research tool for displaying the systematic associations of genetic and epigenetic variations, especially in terms of cytosine methylation onto the genomic region in a specific cell or tissue. A comprehensive study of plant populations to correlate genotype to epigenotype and to phenotype, and also the study of methyl quantitative trait loci (QTL) or epiGWAS, is possible by using high-throughput sequencing methods, which will further accelerate molecular breeding programs for crop improvement.

C. B. Yadav, G. Pandey, M. Muthamilarasan, and M. Prasad (✉)
National Institute of Plant Genome Research (NIPGR), New Delhi, India
e-mail: manoj_prasad@nipgr.ac.in

Graphical Abstract



Keywords Chromatin modification, Crop improvement, DNA methylation, Epigenetics, Epigenomics

Contents

1	Introduction	238
2	Epigenetics	239
2.1	DNA Methylation	239
2.2	Chromatin Modifications	240
2.3	RNA-based Control Mechanisms	242
3	Epigenomics	243
4	Strategies for Genome-wide Epigenetic Profiling for High Resolution of Epigenome	244
4.1	Bisulfite Sequencing	245
4.2	Digestion with Methylation-Sensitive Restriction Enzymes	249
4.3	Chromatin Immunoprecipitation	252
4.4	Small RNA Sequencing for Their Possible Role in Chromatin Modifications	254
5	Epigenomics in Crop Plants	255
6	Conclusion and Outlook	256
	References	256

1 Introduction

In plants, epigenetic regulation of the genome plays a significant role in normal growth and development. DNA methylation, which is a crucial constituent of epigenetic phenomena, controls gene expression during plant growth and development. These epigenetic marks recruited through methylation events are heritable to successive generations. DNA methylation also plays a crucial role in normal plant reproduction and seed development because it is involved in genomic imprinting [1]. In addition to DNA methylation, repeating units of chromatin called nucleosomes are also an important regulator that affect the accessibility of transcription factors and regulators for the expression of genes through various epigenetic mechanisms that involve specific chemical and post-translational modifications of histones. Epimutations, DNA methylation

level, and chromatin remodeling at the genome level may be involved in various kinds of developmental process regulations. These peculiar regulation mechanisms may also cause various kinds of developmental abnormalities, such as sterility, transposon activation, and defects in flowering response pathways [2]. These epigenetic regulators studied at the genome level are called epigenomics. At present, epigenomic studies are possible by microarrays and high-throughput-sequencing technologies that will help in unfolding the complex network of epigenomic regulation and genome activity of plants.

2 Epigenetics

The term “epigenetics” was proposed by Waddington in the 1940s, and he defined it as “causal interactions between genes and their products which bring the phenotype into being.” Later, a more concrete definition of epigenetics was formulated, whereby it is defined as “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence [3].” Epigenetic modifications involve DNA methylation and chromatin modifications. In DNA methylation, the 5' position of cytosine and the N6 position of adenine bases are methylated. The rate of DNA methylation varies in different species with 14% of methylated cytosine reported in *Arabidopsis thaliana*, 4% in *Mus musculus*, 2.3% in *Escherichia coli*, and 0.03% in *Drosophila* [4]. Histone modifications include lysine and arginine methylation, lysine acetylation, ubiquitination, sumoylation, and serine and threonine phosphorylation towards the regulation of gene expression. Epigenetic modifications are species-, tissue-, organelle-, and age-specific, and are involved in various processes like transposon repression, genomic imprinting [5], and stress-associated defense responses.

2.1 DNA Methylation

DNA methylation is an enzyme-catalyzed addition of a methyl (-CH₃) group to the fifth position of cytosine to form 5-methylcytosine. DNA methylation is not only restricted to prokaryotes but also occurs in eukaryotes. Cytosine methylation is common in both animals and plants, whereas adenine methylation is restricted to prokaryotes. In bacteria, the DNA methylation mechanism is slightly peculiar as this helps bacteria in differentiating the host genomic DNA from invading phage DNA, and eventually leads to the cleavage of phage DNA by host restriction enzymes. DNA methylation mechanisms are mostly conserved in eukaryotes such as fungi, plants, and animals. In plants, DNA methylation takes place in three different sequence contexts, viz. CG, CHG, and CHH (H = A, C, or T), catalyzed by DNA methyl transferases. DNA methylation is a reversible, enzyme-mediated modification of bases (Fig. 1). Enzymes responsible for cytosine methylation in plants fall under three distinct categories: First is DNA METHYLTRANSFERASE1 (MET1), a homologue of mammalian methyltransferase (Dnmt1), which is required for maintaining symmetric cytosine methylation (CpG) of the genome. The MET1-silenced plant showed a lack of widespread CpG methylation

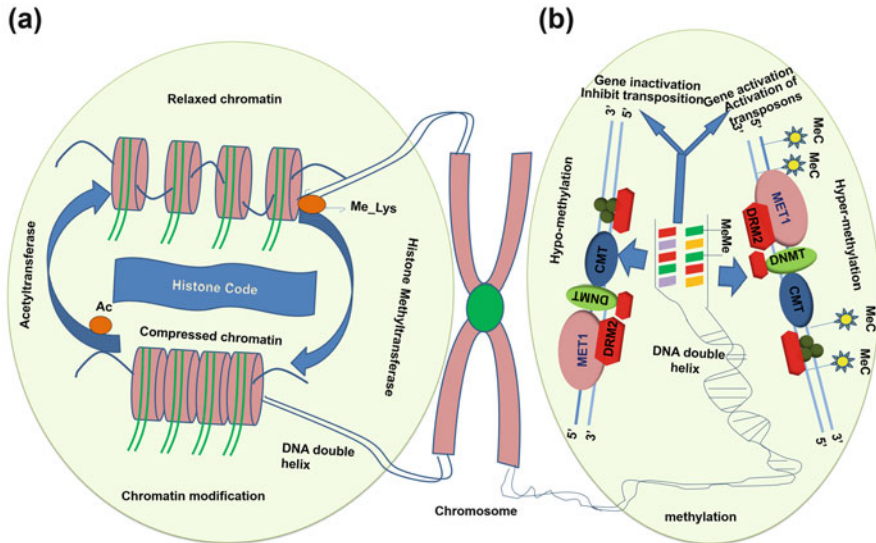


Fig. 1 Diagrammatic representation of chromatin remodeling via histone modification and methylation mechanisms. **(a)** Histone methylation and acetylation alter the gene expression via DNA condensation and relaxation. **(b)** Alterations in gene expression could occur by DNA methylation that involves recruitment of a methyl group to the fifth position of cytosine by CMT3, chromomethylase; DRM3, domain rearrangement methyltransferase; MET1, methyltransferase during DNA methylation events that lead to hypomethylation to activate the gene/transposon or while hypermethylation reduces the activity of the gene/transposon

[6]. Another variant of MET1 showed various phenotypic changes after creating a mutation in the functional gene region. For example, *met1a* mutants in rice showed a normal phenotype, whereas *met1b* mutants display aberrant seed development [7, 8]. In this example, DNA methylation was suddenly decreased at specific regions of the genome such as repetitive centromeric, transposon, and retrotransposon sequences [7, 8]. The second category includes plant-specific CHROMOMETHYLASE3 (CMT3), which recruits a methyl group at the CHG type of sequence, especially at centromeric repeats as well as at transposons [6, 9]. The above-mentioned DNA methyltransferases create a symmetric type of cytosine methylation in the genome [10]. The third category catalyzes asymmetrical cytosine methylation at the CpNpNp site and includes two DNA methyltransferases, DRM1 (DOMAIN REARRANGED METHYLASE1) and DRM2, which are responsible for de novo methylation [11] (Fig. 1).

2.2 Chromatin Modifications

The other epigenetic modification associated with the regulation of gene expression is the covalent post-translational modification of the N-terminal tail of core histone proteins in certain amino acid residues such as lysine, arginine, serine, and threonine.

These modifications will either activate or repress the transcription, depending on the type of histone modification. Histone acetyltransferases (HATs) catalyzed acetylation of H3 and H4 lysine (K) at positions 4, 9, 27, 36, and 73 and is important for the positive regulation of transcription whereas deacetylation catalyzed by histone deacetylase (HDAC) leads to negative regulation. The other form of modification i.e., methylation, affects transcription in a way depending on position and degree of methylation. Lysine and arginine methyltransferases mediate the methylation of histone proteins in lysine (K) and arginine (R) residues, respectively. Active transcription is associated with H3K36me, H3K48me, H3K79me, and trimethylation of histone H3 at lysine 4 (H3K4me3). In contrast, methylation at H3K9, H3K27, H4K20, and H4R3me2 are responsible for chromatin condensation, thus resulting in transcription repression (Fig. 1). Transcriptional activation is also linked to H3 phosphorylation at serine and threonine residues [12].

Understanding the role of histone modification in developmental reprogramming and in response to a range of environmental adversity in plants has been progressed in recent years. Histone modifications are important players in vernalization and photomorphogenesis. H3K4me3 and histone acetylation are responsible for active expression of *Flowering Locus C (FLC)* resulting in late flowering, whereas H3K9me2, H3K27me2, and histone deacetylation reverse this effect by repression of *FLC* in *A. thaliana* [13]. In addition to flowering time regulation, histone modification, particularly H3K9ac, contributes to light-induced activation of HY5 and HYH and their downstream effectors like photosynthesis-related genes such as *photosystem I subunit F (PsaF)* [14]. The promoter and coding region of another photosynthetic gene like *phosphoenolpyruvate carboxylase (Pepc)* in maize showed light-induced acetylation of H4K5 and H3K9 [15]. Moreover, regulation of gibberellin metabolism genes is also associated with light-induced acetylation of H3K27ac and trimethylation of H3K27me [14].

Histone modifications are also important regulatory mechanisms involved in the abiotic stress response. H3K4me3 is a positive regulator of water stress response and it is well illustrated when transcription of *NCED3 (9-cis-epoxycarotenoid dioxygenase)*, which encodes an important ABA biosynthesis enzyme, is reduced in *Arabidopsis* trithorax-like factor *ATX1* mutant (which trimethylates H3K4). This results in a decrease in dehydration tolerance [16]. Histone methylation is important in imparting salinity tolerance by regulating the expression of certain salinity stress-induced transcription factors such as *MYB*, *b-ZIP*, and *AP2/DREB* family members in soybean [17]. Understanding the mechanism behind low temperature tolerance is an important strategy for developing cold-acclimatized plants. In this context, the dynamics of H3K27me3 are correlated with the transcriptional regulation of two cold responsive genes, *COR15A (cold-regulated 15A)* and *ATGOLS3 (galactinol synthase 3)* in *Arabidopsis thaliana*, which showed that levels of H3K27me3 gradually decreased in the promoter region of these two genes upon exposure to cold temperature [18]. Histone modification also plays a significant role in gene regulation under high temperature conditions. Repressive chromatin marks the H3K9me2 level of *Fertilization-Independent Endosperm1 (OsFIE1)* as sensitive to temperature variation during seed development in *O. sativa*. The H3K9me2 level of *OsFIE1* reduces under moderate heat

exposure, leading to increased *OsFIE1* expression [19]. FERTILIZATION-INDEPENDENT SEED (FIS) Polycomb group (PcG) proteins are an evolutionarily conserved class of proteins that are involved in fertilization-independent seed formation and also ensure the stable transmission of developmental decisions. The FIS Polycomb complex alter the target genes by implementing repressive methylation on histone H3 lysine 27. Further, it is also involved in endosperm proliferation and reproductive developments. However, most importantly, the imprinting phenomenon during seed formation and in the endosperm development is controlled by the FIS complex along with DNA methylation [20]. Recently, it was reported that FIE does not have repressive functions in apomictic *Hieracium*, and on down-regulation of FIE, autonomous embryo development is blocked in FIS1:HFIE:RNAi but autonomous endosperm development is capable of occurring. Thus, it was found that maternal FIE is a requirement in an apomictic *Hieracium* [21]. FIE also regulates methyl transferase gene (MET1) expression in ovules as MET1 expression is up-regulated in FIS1:HFIE:RNAi lines.

2.3 RNA-based Control Mechanisms

Small interfering RNAs (siRNAs) lead to de novo DNA methylation in a sequence similarity specific manner at CG, CHG, and CHH sites. It was first discovered by Wassenegger in plants [22] and called RNA-directed DNA methylation (RdDM). The pathway of siRNA biogenesis starts from the generation of double-stranded RNAs (dsRNAs). The source of dsRNAs may be transposable elements, transcribed inverted repeats, or intermediates of viral replication. RdDM is initiated by RNA polymerase IV (POL IV), which generates single-stranded RNA (ssRNA). This ssRNA serves as a template for the generation of dsRNA catalyzed by RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) with the help of the chromatin remodeler CLASSY 1 (CLSY1). This dsRNA is further cleaved by DICER-LIKE 3 (DCL3) into 24-nt siRNA, having 3' overhangs, which are subsequently methylated by HUA-ENHANCER 1 (HEN1). A single strand of this methylated siRNA is then loaded on ARGONAUTE 4 (AGO4) and forms an RNA-induced silencing complex (RISC)-AGO4 complex, which then guides the methylation of homologous loci. Simultaneously, at the target site, PolV transcribes long non-coding RNA (lncRNA) with the help of DDR complex (DRD1 (DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1), DMS3 (DEFECTIVE IN MERI-STEM SILENCING 3), RDM1 (REQUIRED FOR DNA METHYLATION 1), and DMS4. Previous reports proposed that DDR helps to unwind the DNA for transcription [23]. The (RISC)-AGO4 complex associates with PolV by base pairing of siRNA with lncRNA and this association is stabilized by the interaction of AGO4 with subunits of PolV-NUCLEAR RNA POLYMERASE E1 (NRPE1) and NRPE2 along with KTF1 (KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR 1). RDM1 binds with AGO4 and DRM2 (de novo methyltransferase) leading to cytosine methylation at the target site [24] (Fig. 2).

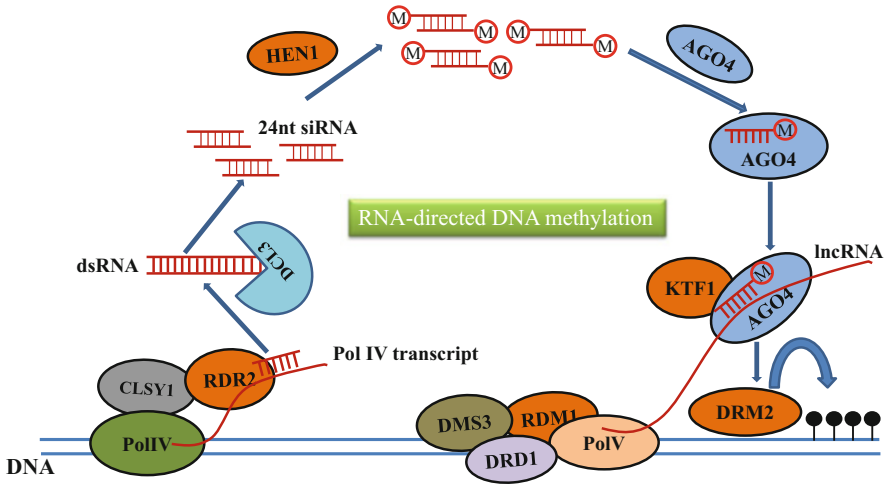


Fig. 2 Diagrammatic representation of RNA-directed DNA methylation (RdDM) pathways. Initially, RDR2 physically interacts with Pol IV and results in the conversion of Pol IV transcripts into double-stranded RNA (dsRNA) with the help of the chromatin remodeler CLSY1. The dsRNAs are further processed into 24-nucleotide (nt) siRNAs by the action of DCL3, and the guide strand is loaded onto AGO4, which then enters the Pol V-mediated pathway of de novo DNA methylation

3 Epigenomics

Epigenomics are the mechanisms that contribute to regulation of the genome through various processes of epigenetics. For example, genome regulation by the action of DNA methylation, histone modifications, and lncRNA expression into a particular tissue, or even one particular cell type or organ. Unlike genomics, epigenomics is a dynamic process that can be influenced by environmental factors such as biotic and abiotic stress in a particular tissue or organ. Aberrant changes in the genome through various epigenetic events lead to distortion in morphological, developmental stages and results in plant abnormalities in the form of plant diseases. Nowadays, the most challenging task is to understand how the epigenome contributes to gene regulation, which in turn will give us greater insight into plant development. High-resolution epigenome mapping could be done through genome-wide analysis of DNA methylation, histone modifications, and siRNAs in correlation with chromatin accessibility and finally to mRNA transcription in plants. In the era of next-generation sequencing, genome-wide epigenetic changes could be quantified in the genome through various high-throughput techniques such as DNA methylation, histone modification, chromatin remodeling, and regulatory noncoding RNAs. A variety of next generation sequencing platforms are available for identifying the epiallelic variation in the genome.

The three most popular distinctive sequencing platforms: 454 Genome Analyzer FLX (Roche), the HiSeq (Illumina), and the 5500xl SOLiD System (Life

Table 1 Summary of the most popular next-generation sequencing platforms

Method	Accuracy (%)	Read length (bp)	Reads per run	Advantages	Disadvantages
Pyrosequencing (454)/Roche	99.90	700	1 million	Long read size. Fast	Runs are expensive
Illumina/Solexa	99.90	2 × 100	3 billion	Potential for high sequence yield	Equipment can be very expensive
Life technologies (SOLiD)	99.90	200–400	4 billion	Low cost per base	Has issues sequencing palindromic sequences
Ion Torrent sequencing	98	40,000%	Up to 80 million	Less expensive equipment	Homopolymer errors

Technologies) have since been developed for the generation of large-scale reads and higher throughput (Table 1). The different sequencing platforms have particular advantages that are utilized on the basis of the required scientific goals during research. For epigenomic studies, the number of sequenced reads to consider the best-read depth and -read length for genome coverage are important key factors to choose the best suitable sequencing platform for every experiment. HiSeq and SOLiD are the best suited sequencing platform for epigenomic studies including transcriptomics, sRNA analysis, ChIP-Seq, and DNA methylome analysis because a large number of reads can be obtained using these platforms. 454 Genome Analyzer FLX sequencer produces longer reads, hence it is presently best suited for de novo genome and transcriptome assemblies. Detailed technical background regarding different NGS platforms with respect to sequence read generation and sequencing reactions are extensively described by Buermans et al. [25].

4 Strategies for Genome-wide Epigenetic Profiling for High Resolution of Epigenome

During the last few decades, it has been revealed that gene expression regulation or repression through DNA methylation entails several steps that suppress the gene response within partial pathways; however, it is still unclear whether gene body and intergenic region methylation could play a crucial role in gene expression. Nowadays it is well established that the promoter region of gene is a strong key factor for gene silencing. Generally, promoter regions are the CpG-rich region as compared to other parts of the genome and these CpG rich nucleotides are methylated. In plants, several environmental factors that induce a hypermethylation state in the promoter regions, which has CpG-islands and concern genes, becomes inactivated. Therefore, differentially methylated regions (DMRs) in the plant genome could be quantified even in response to various stress conditions. In genome DNA, methylation levels increase through the action of DNA methyl transferases (DNMTs), which provide an

epigenetic mark for the recognition of methyl-DNA binding proteins. Ultimately, other multi-protein complexes possessing chromatin-modifying activity, for instance histone-deacetylases (HDACs), and chromatin remodelers are recruited in the genome to establish a repressive chromatin configuration.

Techniques to detect the DNA methylation and chromatin modification patterns in the genome can be classified into following categories: (1) bisulphite conversion, (2) digestion with methylation sensitive restriction enzymes, (3) chromatin immunoprecipitation mediated with high throughput sequencing (chip-seq), and (4) small RNA-mediated methylation (Table 2). The above-mentioned techniques differ in their principle of distinguishing methylated from unmethylated DNA.

4.1 Bisulfite Sequencing

The bisulfite conversion method involves treating genomic DNA with sodium bisulfite, which leads to the conversion of unmethylated cytosine into uracils, whereas methylated cytosine residues remain unaffected during the treatment. This converted DNA can be amplified with PCR by using sequence-specific primers and finally the methylation status of the DNA can be revealed. Thus, bisulfite treatment could be a promising tool to detect specific modifications in a DNA sequence that relies on the methylation status of individual cytosine residues, providing single-nucleotide resolution information on the methylation status of a DNA sequence. After that, various bioinformatic analyses can be carried out on the converted sequence to recover the information for nucleotide resolution. The advancement of sequencing techniques, especially Illumina sequencing, i.e., sequencing by synthesis technology, provides us with an opportunity to sequence the entire cytosine methylome of a genome at single-base resolution (methylC-seq). This single-base methylome map provides us with earlier undetected DNA methylation, facilitates the determination of context as well as of the level of methylation at each site, and the effect on the state of DNA methylation influenced by nearby sequence composition. This whole genome bisulfite sequencing would also provide the methylation level in promoters, UTRs, and other protein-coding regions of the gene. Small RNA and transcriptome sequencing, and their direct association between abundance of sRNAs and DNA methylation, can be quantified with bisulfite sequencing. Strand-specific mRNA-sequencing revealed altered transcript abundance of certain genomic regions such as transposons, intergenic regions, and gene changes in the transcript abundance of hundreds of genes, transposons, and unannotated intergenic transcripts upon changing their DNA methylation state. In brief, these complete and well-integrated data sets divulge earlier unexplored subsets of the epigenome and help in understanding the intricate relationship between DNA methylation and transcription.

Nowadays many researchers are using bisulfite sequencing-based methods for studying methylation across entire genomes of plant, populations, and plant species. Whole genome bisulfite sequencing (WGBS) is possible for identifying the genetic basis for phenotypic variation within large populations either in a segregating

Table 2 Epigenomics studies for methylation quantification in diverse plant species

Methodology	Plant	Nature of study	References
ChIP (MeDIP)	Maize	Maize methylome profiling in response to abiotic stress	Eichten and Springer [26]
Chip (MeDIP)	Arabidopsis	The progeny of <i>Arabidopsis thaliana</i> plants exposed to salt exhibit changes in DNA methylation, histone modifications, and gene expression	Bilichak et al. [27]
ChIP	Maize	Epigenetic regulation of the cell wall related genes in response to salt stress	Li et al. [28]
ChIP	Arabidopsis	Decrease in H3K27me3	Kwon et al. [18]
ChIP	Maize	Increase in H3K9ac and H4K5ac in promoter of cell cycle genes	Zhao et al. [29]
ChIP	Rice	Transcriptional activation of <i>OsDREB1b</i> by hyperacetylation H3	Roy et al. [30]
ChIP	Maize	Selectively suppresses the cold-induced transcription of the <i>ZmDREB1</i> gene in maize	Hu et al. [31]
ChIP-qPCR analysis	Arabidopsis	Transgenerational phenotypic and epigenetic changes in response to heat stress in <i>Arabidopsis thaliana</i>	Migicovsky et al. [32]
ChIP seq	Arabidopsis	Increased sodium transporter gene in response to salt stress	Sani et al. [33]
ChIP seq	<i>Physcomitrella patens</i>	Changes in H3K4me3, H3K27Ac, and H3K9Ac during drought stress	Widiez et al. [34]
ChIP seq	Rice	Genome-wide profiling of histone H3K4-trimethylation and gene expression in rice under drought stress	Zong et al. [35]
Bisulfite sequencing	Arabidopsis	Reference-guided assembly of four diverse <i>Arabidopsis thaliana</i> genomes	Schneeberger et al. [36]
Bisulfite sequencing	Tomato	Epigenetic marks in an adaptive water stress-responsive gene in tomato roots under normal and drought conditions	González et al. [37]
Bisulfite sequencing	Maize	CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize	Gent et al. [38]
Bisulfite sequencing	Rice	Plants regenerated from tissue culture contain stable epigenome changes in rice	Stroud et al. [39]
Bisulfite sequencing	Soybean	Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population	Schmitz et al. [40]
Bisulfite sequencing	Brachypodium	Gene body methylation is conserved between plant orthologs and is of evolutionary consequence	Takuno et al. [41]
Bisulfite sequencing	<i>Capsella rubella</i>	Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization	Seymour et al. [42]
MSAP	Rice	Salt stress-induced variation in DNA methylation pattern and its influence on gene expression in contrasting rice genotypes	Karan et al. [43]

(continued)

Table 2 (continued)

Methodology	Plant	Nature of study	References
MSAP	Rice	DNA methylation changes detected by methylation-sensitive amplified polymorphism in two contrasting rice genotypes under salt stress	Wang et al. [44, 45]
MSAP	Wheat	DNA-methylation changes induced by salt stress in wheat <i>Triticum aestivum</i>	Zhong et al. [46]
MSAP	Maize	Analysis of DNA methylation of maize in response to osmotic and salt stress based on methylation-sensitive amplified polymorphism	Tan et al. [47]
MSAP	Brassica	Use of MSAP markers to analyze the effects of salt stress on DNA methylation in rapeseed (<i>Brassica napus</i> var. <i>oleifera</i>)	Marconi et al. [48]
MSAP	Rice	Drought-induced site-specific DNA methylation and its association with drought tolerance in rice (<i>Oryza sativa</i> L.)	Wang et al. [44, 45]
MSAP	Rice	Transgenerational variations in DNA methylation induced by drought stress in two rice varieties with distinguished difference to drought resistance	Zheng et al. [49]
MSAP	Rice	Epigenetic responses to drought stress in rice (<i>Oryza sativa</i> L.)	Gayacharan and Joel [50]
MSAP	Brassica	Comparison of the heat stress induced variations in DNA methylation between heat-tolerant and heat-sensitive rapeseed seedlings	Gao et al. [51]
MSAP	Grapevine	Dynamics and reversibility of the DNA methylation landscape of grapevine plants (<i>Vitis vinifera</i>) stressed by in vitro cultivation and thermotherapy	Baranek et al. [52]
MSAP	Rice	Transgenerational inheritance of modified DNA methylation patterns and enhanced tolerance induced by heavy metal stress in rice (<i>Oryza sativa</i> L.)	Ou et al. [53]
MSAP	Poplar	Epigenetic control of heavy metal stress response in mycorrhizal versus non-mycorrhizal poplar plants	Cicatelli et al. [54]
MSAP	<i>Posidonia oceanica</i>	In <i>Posidonia oceanica</i> cadmium induces changes in DNA methylation and chromatin patterning	Greco et al. [55]
MSAP	Radish	Analysis of genomic DNA methylation level in radish under cadmium stress by methylation-sensitive amplified polymorphism technique	Yang et al. [56]
MSAP	Brassica	The protective role of selenium in rape seedlings subjected to cadmium stress	Filek et al. [57]
MSAP	Chickpea	DNA methylation and physio-biochemical analysis of chickpea in response to cold stress	Rakei et al. [58]

population or in natural germplasm lines. Recently, a large number of research groups are focusing on WGBS in a number of plants to generate the methylomes maps, ranging from model plants like *A. thaliana* [40, 59, 60] to economically important crops like *Z. mays* [28, 61, 62]. This approach could be important for the study of evolutionary epigenomics and comparative epigenomics to understand both the variable and also the invariable portions of epigenomes by profiling DNA methylomes, histone tail modifications, and RNAs from a variety of flowering plant species. The use of WGBS together with de novo transcript assemblies has provided an opportunity to monitor the changes in methylation of gene bodies among species [63], but does not provide a full view of changes in the patterns of context-specific methylation at different types of genomic regions [42]. Bisulfite sequencing (BS-seq) was done by Takuno et al. [63] by taking two different tissues (leaves and immature floral buds) of *Brachypodium distachyon* to see the comparative methylation pattern in *B. distachyon* tissues and also among *B. distachyon* and rice (*Oryza sativa* ssp. japonica) [64]. Cytosine DNA methylation through whole-genome bisulfite sequencing was carried out by Lister et al. [60] in *Arabidopsis*, in soybean by Schmitz et al. [40], in tomato by Zhong et al. [65], and in maize by Gent et al. [38]. Stroud et al. [39] investigated the effect of DNA methylation through bisulfite sequencing in rice plant regenerated through tissue culture and compared the single-base resolution maps of DNA methylation of transformed, regenerated rice lines with non-transformed, regenerated rice lines. They found that tissue culture practice induces stable changes in DNA methylation in regenerated plants, resulting in ectopic losses of DNA methylation in regenerated lines.

An alternative and most effective cost-reduced bisulfite sequencing approach called the reduced-representation bisulfite sequencing (RRBS) approach has been developed by Meissner et al. [66] to investigate the mammalian methylome. The RRBS method involves MspI restriction enzyme digestion of genome followed by bisulfite conversion and subsequently next-generation sequencing to analyze methylation patterns of specific fragments. RRBS-based protocols are more economic since these methods rely on the enrichment of CpG-rich regions in close proximity to the recognition sequence of restriction enzymes; however, these protocols might show lack of coverage at intergenic and distal regulatory elements that are relatively less studied. Methylation quantification with RRBS has been widely used in the profiling of plant methylomes on large-scale samples for the demonstration of epigenome-wide association studies (EWAS). Schmitz et al. [40] performed RRBS in 83 soybean recombinant inbred lines (RILs) and their parents, to identify the patterns and heritability of methylation variants for understanding how methylation variants contribute to phenotypic variation. The RRBS method was also applied by [67] in *Brassica rapa* to decipher the role of epigenetic variation and it was suggested that these epigenetic variations could play a strong role in polyploid genome evolution and also could be an alternative mechanism for duplicate gene loss.

Bisulfite sequencing data could provide the methylation state of cytosine residues at a single-base resolution. However, a systematic analysis of sequencing data is required for statistical evaluation of methylation at all possible sites in the complete

genomic region. With the advancement of sequencing and availability of methylome data, publicly accessible computational tools are required to analyze the data. For example, a web-based tool Cytosine Methylation Analysis Tool (CyMATE), could be used for aligning the bisulfite-converted sequence with a reference sequence to get the methylation pattern at CG, CHG, and CHH (H = A, C, or T) sites, in each sequence and at the single-base position. Similarly, another bioinformatics tool for bisulfite sequencing-data evaluation is Kismeth, which is used to find out the cytosine methylation in different sequence contexts (CG, CHG, and CHH). This tool can also be used for designing bisulfite primers as well as for the analysis of the bisulfite sequencing results.

Besides web-based tools, various standalone computational tools are available that help in the quantitative assessment of bisulfite sequencing data obtained from methylation changes occurring in plants. An example is a computational pipeline (methylKit) developed by Akalin et al. [68], which is a multi-threaded R package that can quickly and simultaneously analyze and characterize data from a set of methylation experiments. It is a user-friendly tool as it can use a text file as well as an alignment file to read DNA methylation information. Comparative differential methylated regions among individuals can be carried out with this tool. MethylKit would also carry out the categorization of the sample, as well as the annotation and visualization of DNA methylation events. Similarly, another methylome analysis pipeline (Methy-Pipe) was developed by Jiang et al. [69], which is an efficient and integrative bioinformatics software package for methylation data analysis along with downstream analysis. A flexible and time-efficient tool (Bismark) for the analysis of bisulfite sequencing data was developed by Krueger and Andrews [70]. This provides a snapshot of a cell's epigenomic state by figuring out its cytosine methylation at a single-base resolution for the complete genome. This tool can be used to map the reads and methylation, using only a single step to distinguish the methylated cytosines in the CG, CHG, and CHH context, and it facilitates the analysis and interpretation of researchers' methylation data. There are some software packages that are designed for bisulfite sequencing read alignment only, for example see Chen et al. [71], Krueger et al. [70], Lim et al. [72], Xi et al. [73], whereas downstream analysis requires specific software packages for visualization and comparative analysis [74, 75] (Table 3).

4.2 Digestion with Methylation-Sensitive Restriction Enzymes

A variety of techniques have been developed for quantifying DNA methylation without any prior information of the DNA sequence, including methylation-sensitive amplified polymorphism (MSAP). The MSAP method was established to identify the cytosine methylation pattern in the genomes. This method involves the use of two methylation-sensitive isoschizomers viz. Hpa II and Msp I, which differ in their sensitivity to the methylation status of the same recognition sequences (5'-CCGG-3'). HpaII cannot cut when cytosine is fully methylated (both strands methylated) but

Table 3 Tools for methylation quantification

Name	Application	Web-link	Reference
<i>Bisulfite</i>			
PEAR	Merging raw Illumina paired-end reads (RRBS)	http://www.exelixis-lab.org/web/software/pear	Zhang et al. [76]
R/ Bioconductor package	Identification of differentially methylated regions	–	–
<i>DMRcate</i>	Differentially methylated regions (DMRs)	https://www.bioconductor.org/packages/release/bioc/src/contrib/DMRcate_1.8.5.tar.gz	Peters et al. [77]
MOABS	Analysis of large scale base-resolution DNA methylation	https://s3.amazonaws.com/deqiangsun/software/moabs/moabs-v1.3.2	–
DMAP	DMRs	http://biochem.otago.ac.nz/assets/software/meth_progs_dist.tar.gz	–
Methpipe	WGBS and RRBS	http://smithlabresearch.org/downloads/methpipe-3.4.2.tar.bz2	Song et al. [78]
Minfi	Tools for analyzing and visualizing Illumina's 450k array data	http://bioconductor.org/packages/release/bioc/src/contrib/minfi_1.18.2.tar.gz	Aryee et al. [79]
BEAT	BS-Seq Epimutation Analysis Toolkit	https://www.bioconductor.org/packages/release/bioc/src/contrib/BEAT_1.10.0.tar.gz	Akman [80]
Bismarch	A flexible aligner and methylation caller for Bisulfite-Seq applications	http://www.bioinformatics.bbsrc.ac.uk/projects/bismark/	Krueger et al. [70]
Mehtylkit	A comprehensive R package for the analysis of genome-wide DNA methylation profiles	http://code.google.com/p/methylkit	Akalin et al. [68]
<i>Epigenomics</i>			
coMET	Visualization of EWAS results in a genomic region	http://epigen.kcl.ac.uk/comet	Martin et al. [81]
Repitools	Analysis of enrichment-based epigenomic data	http://bioconductor.org/packages/release/bioc/src/contrib/Repitools_1.18.0.tar.gz	Statham et al. [82]
methylPipe and compEpiTools	Epigenomics	https://doi.org/10.1186/s12859-015-0742-6	Kishore et al. [83]
RnBeads	Comprehensive analysis of DNA methylation data	http://rnbbeads.mpi-inf.mpg.de/installation.php	–
ALEA	Computational toolbox for allele-specific (AS) epigenomics analysis	ftp://ftp.bcgsc.ca/supplementary/ALEA/files/ALEA-userguide.pdf	–

(continued)

Table 3 (continued)

Name	Application	Web-link	Reference
Epigenomix	Analysis of RNA-seq or microarray based gene transcription and histone modification data obtained by ChIP-seq	http://epigenie.com/epigenetic-tools-and-databases/	Klein et al. [84]
MACS	Analysis of ChIP-Seq	–	Zhang et al. [85]
EaSeq	Visualization of ChIP-sequencing data	http://epigenie.com/epigenetic-tools-and-databases/	–
MMDiff	Analysis of ChIP-Seq data sets	http://homepages.inf.ed.ac.uk/gschweik/MMDiff.html	Schweikert [86]
ODIN	ChIP-seq signals with differential peaks	http://costalab.org/wp-content/uploads/2014/04/ODIN-sim.tar.gz	–

cleaves when external cytosine is hemimethylated; in contrast, MspI cuts hemi- or fully methylated C5mCGG but not 5mCCGG. In this way, based on restriction sites, the locus-specific discrimination between methylated and unmethylated DNA sequences can be identified. A number of research groups have investigated the methylation/demethylation events in plants using the MSAP technique. For example, the change in methylation pattern has been investigated by Zheng et al. [49] in response to drought, in salt stress conditions by Karan et al. [43], in heat conditions by Gao et al. [51], in heavy metals by Ou et al. [53], and in aluminum by Choi and Sano [87].

Several research groups have utilized the MSAP technique to quantify the differentially methylated regions for important agronomic traits by comparing the methylation pattern in contrasting cultivars. For example, Marconi et al. [48] reported that methylation levels were greatly reduced in a tolerant cultivar (Exagone) of *Brassica* in response to salinity stress, whereas the sudden enhancement of methylation was recorded in a susceptible *Brassica* cultivar (Toccatà) under salinity stress. Similar kinds of methylation patterns in tolerant and susceptible cultivars were also observed in foxtail millet under salt stress conditions. However, in some cases, methylation events are highly dependent upon specific types of tissue and genotype rather than the tolerance or susceptibility of plants. For example, Karan et al. [43] investigated rice cultivars and found that the tolerant cultivar (Pokkali and IR29) and Nipponbare, which is sensitive to salinity stress, showed tissue and genotype-specific methylation/demethylation events under salt stress, which was totally irrespective of the tolerance and susceptibility of the plant [43]. Similarly, in *Vitis vinifera*, the authors have characterized for multiple stresses and found that the Sangiovese cultivar showed a sensitive response for photo inhibition manifested as incomplete damage to plant leaves, whereas the Montepulciano cultivar does not show any such response. These tolerant and susceptible cultivars were screened for differential methylation patterns using MSAP and many differential methylated regions were found in both the cultivars during drought stress conditions [88].

4.3 Chromatin Immunoprecipitation

Accessibility of nucleosome to regulatory proteins is an important key regulator for the expression of genes. Chromatin modifiers as well as histone proteins are key players in transcriptional regulation through changing the compactness of DNA through nucleosome rearrangements. These dynamic alterations in chromatin are also denoted as the epigenome, which are different for distinct tissue types, developmental stages, and disease states, and also dynamic in response to environmental changes. These kinds of variations in chromatin state or epigenomic phenomena can be quantified with recent high-throughput technologies at the genome-wide level. Currently, a popular technique is the chromatin immunoprecipitation (ChIP) assay, which is used to study the epigenome. ChIP assays or ChIP sequencing (ChIP-Seq) is a useful method for discovering genome-wide modifying positions in the chromatin complex containing transcription factors and other proteins. This could be an important approach for studying the histone or other protein–DNA interactions in a particular tissue type, in cells of different developmental stages, or in cells altered by various environmental factors. This will also display a deep insight into gene regulation mechanisms during exposure to diverse environmental stresses and biological pathways involved in plant development and growth. Using this technique, a genome-wide interaction between proteins and nucleic acids could be examined. The ChIP method involves crosslinking, isolation, and fragmentation of chromatin followed by capturing of the protein–DNA complexes by using antibodies against the histone or transcription factor under study. The immunoprecipitated protein–DNA complex are reverse crosslinked, DNA is then purified for further analysis either by hybridization to microarrays, i.e. ChIP-chip, or by high-throughput sequencing (ChIP-seq).

In the ChIP–chip technique, plant materials containing histones and DNA called nucleosome complex are crosslinked with formaldehyde followed by extraction and fragmentation of chromatin. Finally, these sheared fragments are allowed to undergo chromatin immunoprecipitation (ChIP) with modification-specific antibodies. The enrichment process could be performed with PCR amplification to obtain adequate DNA that is then denatured to get the single-strand DNA (ssDNA). ssDNA fragments are subjected to labelling with fluorescent tags to differentiate the samples. Finally, these labeled fragments are used for hybridization to the target single-stranded sequences spotted on the DNA microarray surface representing the genomic regions of interest. The complementary fragments of labelled fragments will hybridize on the target sequences on the chip array to form double-stranded DNA fragments followed by illumination with a fluorescent light. The fluorescence signals generated from the array are normalized with control signals, and statistical tests are applied to find out the methylated region. The existing coordinates of microarray probes can then be mapped onto the reference genome to find their physical positions. This method was found to be useful in studying histone modifications linked with C4 photosynthesis in maize [15, 89] and systemic immunity in *Arabidopsis* [90]. Several studies in different plant systems have been performed using immunoprecipitation combined with microarray hybridization for epigenomic

studies. ChIP on ChIP was performed by Eichten and Springer [26] in maize genomes to estimate the DNA methylation under environmental stress such as cold, heat, and UV stress. A comparison of the DNA methylation pattern suggested a low-rate of putative variation present between control and cold, heat and UV-stressed plants. Similarly, Zhang et al. [91] investigated the distribution pattern of mono-, di-, and trimethylated H3K4 on a genome-wide scale in *Arabidopsis thaliana* seedlings. They used chromatin immunoprecipitation in combination with high-resolution whole-genome tiling microarrays (ChIP-chip). They found that all three methylation patterns showed different distribution patterns in the *Arabidopsis* genome. For example, promoters as well as 5' genic regions were mainly occupied with H3K4me2 and H3K4me3 types of modification; in contrast H3K4me1 was found to be distributed in the transcribed regions. In rice, ChIP-on-ChIP analysis showed that the gene expressions were relatively low when H3K4me was increased, and decreased for genes with high expression levels [35].

However, ChIP-Seq is the technique in which chromatin immunoprecipitation is followed by next-generation sequencing techniques. It has emerged as one of the most interesting and leading technologies for epigenetic study on a genome-wide scale as it relies on the combination of ChIP with next-gen sequencing. The first step is crosslinking the DNA binding protein with the DNA strand, followed by shearing of the DNA along with bounded proteins to obtain the small fragments. These fragments are subjected to immunoprecipitation by using antibodies specific for particular histone modification and finally enriched modified chromatin will be obtained by reverse crosslinking the DNA-protein complex. Again, the ChIP DNA ends are repaired and ligated to a pair of adaptors, followed by PCR amplification using primers compatible with the sequencing platform. Illumina platform or other next-gen sequencing techniques could be used for ChIP library sequencing. Afterwards, the raw data obtained is subjected to processing by the Illumina base-calling pipeline. The large scale of sequence reads that correspond to the immunoprecipitated fragments resulting after sequencing could be mapped onto the reference genome to get their physical positions on the genome. These mapped positions are the classified genomic locations of DNA-binding proteins such as DNA-binding enzymes, transcription factors (TFs), modified histones, chaperones, and nucleosomes, thus revealing the importance of these protein-DNA interactions in gene expression and other cellular processes. A wide range of plant genomes has been scanned for histone alteration mark using Chip seq. In *Arabidopsis thaliana*, a genome-wide distribution of histone H3K4me1, H3K4me2, and H3K4me3 were performed by van Dijk et al. [92] using ChIP-Seq during watered and dehydration stress conditions. They found that one or more of the H3K4 methylation marks are central to ~90% of annotated genes. Widiez et al. [34] have performed genome-wide studies in *Physcomitrella patens* for the mapping of five histone modifications (H3K4me3, H3K27me3, H3K27Ac, H3K9Ac, and H3K9me2) using ChIP-seq on the SOLiD platform and they found that H3K4me3, H3K27Ac, and H3K9Ac, which are activating marks, showed significant changes during early developmental stages in response to drought stress, whereas changes to H3K27me3 are mostly observed for genes differentially expressed at the time of development. Genome-wide histone

modification profiling by chip-seq in moss showed H3K27me3 modification, which plays a crucial role in developmental transitions [93, 94].

The use of NGS for sequencing of ChIP fragments is a potentially strong strategy for providing the relatively high-resolution, low-noise, and high-genomic coverage compared with ChIP-on-chip assays. The resolution of ChIP-on-chip strictly depends on the compactness of, as well as the size of, the chromatin fragments that are used for ChIP and the probes on the array, whereas the resolution of ChIP-Seq depends on sheering of chromatin fragments for generation of equal size fragments, as well as the depth of reads during sequencing. As for the cost to achieve nucleosome resolution in plant genomes, ChIP-Seq is less expensive than ChIP-on-chip, given the current cost of whole-genome tiling arrays.

4.4 Small RNA Sequencing for Their Possible Role in Chromatin Modifications

Small interfering RNA (siRNA) also plays an important role by targeting chromatin to regions of sequence similarity in the genome. These siRNAs typically guide sequence-specific DNA and histone methylation known as RNA-directed DNA methylation (RdDM) and form heterochromatin leading to transcriptional gene silencing [2, 95]. For sRNA-mediated epigenomic study, the most popular method is based on size selection of total RNA population from diverse genotypes, tissue types, mutants, and accessions or subspecies. In this, total RNA is isolated from required samples followed by size fractionation. Subsequently, ligation steps are performed to add DNA adaptors at both the ends of the sRNAs, which act as primer binding sites during reverse transcription and PCR amplification. Finally, genome-wide large-scale reads can be obtained after sequencing using NGS approach.

Recent findings also point to a role for small RNAs derived from transposons to specific regions to regulate expression of genes related to female gametophyte developments [96]. Recent evidence has also indicated the involvement of retrotransposons in tissue-specific silencing leading to heterochromatin formation [96]. It was also shown that the Argonaute 9 (AGO9) gene belonging to the small RNA pathway is indicative of a significant proportion of long terminal repeat retrotransposons (LTRs) in the ovule and that its predominant TE targets are located in the pericentromeric regions of all five chromosomes of *Arabidopsis*. This suggests a link between the AGO9-dependent sRNA pathway and heterochromatin formation during megagametophyte formation. Thus, an understanding of epigenetic regulation during reproductive developments in plants helps researchers to target suitable retrotransposons for the regulation of phenotypic variations. Thus, the chromatin environment can be manipulated using siRNA/miRNA to make certain regions of the genome more or less susceptible to transcription.

5 Epigenomics in Crop Plants

Being sessile in nature, plants adapt to environmental changes through various physiological or developmental adjustments by altering the chromatin structure. This affects several processes, like floral development, flowering time, imprinting, and environmental stress (both biotic and abiotic) responses in plants. Chromatin changes include the process of histone modifications, DNA methylation, and small RNA-mediated silencing to regulate gene expression. Recent advances in sequencing technologies provide us the opportunity to harness the role of epigenetic and epigenomic studies in understanding the regulation of gene expression on perception of developmental and environmental stimuli for crop improvements. Changes in the epigenetic state, chromatin modifications, or DNA methylation under the influence of environmental cues, such as temperature, light, hypoxia, drought, salt stress, and pathogen response, is evident from several studies. Several recent epigenetic studies have also been demonstrated to identify the epimark in the genome towards the regulation of agronomic traits in plants.

Schmitz et al. [40] demonstrated an epigenomic study in soybean recombinant inbred lines (RILs) along with their parents and stated that the majority of methylation variants adhere to Mendelian modes of inheritance but also demonstrate rare examples of epigenetic variation that do not follow the standard laws of inheritance. The interconnection between methylation, gene expression, and genetic variation could be inferred through population epigenomic approaches, which integrate the epigenetic data with expression data derived from transcriptome sequencing or genomic data generated through resequencing [97]. Implementation of population epigenomic approaches to natural and novel experimental populations will unravel the role and effect of DNA methylation in inducing phenotypic variation. These epigenetic variations in segregating population such RILs could be used as tools for the identification of quantitative traits and their associated morphological traits, based on epigenetic variations at particular loci. In plant genomes, most of the variation could arise due to the transposition of transposable elements, which is equally responsible for maintenance of selection pressure. Epigenetic mechanisms are also involved in conferring stress adaptation by inducing post-translational modifications to the N-terminal region of nucleosome core complex histones via acetylation, phosphorylation, ubiquitination, and sumoylation [98, 99]. With the *Arabidopsis* WD-40 protein gene, HOS15 plays a role in histone deacetylation, and this protein is also important for the repression of genes responsible for acclimation and tolerance to cold stress, as HOS15 mutants are hypersensitive to cold stress [100]. Phosphorylation of histone H3, S10, and acetylation of histone H4, is correlated with increased abundance of salt tolerance transcripts in tobacco and *Arabidopsis* [101].

6 Conclusion and Outlook

The mechanism as well as the importance of epigenetic inheritance in plants is now well explored. There is indeed the possibility for a better understanding of epigenetics to facilitate novel and better approaches to crop improvement. The advancement of technologies for rapid and efficient profiling of both genotype and epigenotype will contribute many resources for dissecting the role of epigenetics in imparting variation to important phenotypes and responses to environmental signals. These epigenetic signatures in response to environmental factors that are involved in cell-fate determination, development and cellular proliferations by gene activity modification via histone modifications, DNA methylation, or gene silencing by small RNAs, could be specified with various epigenetic technologies. In view of this, involvement of DNA methylation and small RNA pathways could be identified during plant growth and development by mutating the alleles of genes that will lead to the development of an aberrant phenotype. Thus, an illustration on epigenetic regulation during plant growth and development helps researchers to target suitable genes/transcription factors or genomic regions for the induction of desired phenotypes for further crop improvement programs.

Acknowledgements C.B.Y. acknowledges the Science and Engineering Research Board, Department of Science and Technology, Govt. of India, India for providing a Young Scientist Research Grant (File No. YSS/2015/000287). Ms. Garima Pandey and Mr. Mehanathan Muthamilarasan thank the University Grants Commission for Research Fellowship.

References

1. Grossniklaus U (2001) From sexuality to apomixis: molecular and genetic approaches. In: Savidan Y, Carman J, Dresselhaus T (eds) *Advances in apomixis research*. CIMMYT Press, Mexico City, pp 168–211
2. Lippman Z, Gendrel A, Black M, Vaughn M, Dedhia N, McComble R, Lavine K, Mittal V, May B, Kasschau K, Carrington J, Doerge R, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430:471–476
3. Jablonka E, Lamb MJ (2002) The changing concept of epigenetics. *Ann N Y Acad Sci* 981:82–96
4. Capuano F, Mulleder M, Kok R, Blom HJ, Ralser M (2014) Cytosine DNA methylation is found in *Drosophila melanogaster* but absent in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and other yeast species. *Anal Chem* 86:3697–3702
5. Gehring M (2013) Genomic imprinting: insights from plants. *Annu Rev Genet* 47:187–208
6. Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, Jacobsen SE (2001) Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* 292:2077–2080
7. Hu L, Li N, Xu C, Zhong S, Lin X, Yang J et al (2014) Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. *Proc Natl Acad Sci U S A* 111:10642–10647

8. Yamauchi T, Johzuka-Hisatomi Y, Terada R, Nakamura I, Iida S (2014) The MET1b gene encoding a maintenance DNA methyltransferase is indispensable for normal development in rice. *Plant Mol Biol* 85:219–232
9. Tompa R, McCallum CM, Delrow J, Henikoff JG, van Steensel B, Henikoff S (2002) Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. *Curr Biol* 12:65–68
10. Chan S, Henderson IR, Jacobsen SE (2005) Gardening the genome: DNA methylation in *Arabidopsis thaliana*. *Nat Rev Genet* 6:351–360
11. Ramsahoye BH, Biniszkiwicz D, Lyko F, Clrk V, Bird AP, Jaenisch R (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A* 97:5237–5242
12. Li GF, Bishop KJ, Hall TC (2001) De novo activation of the beta-phaseolin promoter by phosphatase or protein synthesis inhibitors. *J Biol Chem* 276:2062–2068
13. He Y (2009) Control of the transition to flowering by chromatin modifications. *Mol Plant* 2:554–564
14. Charron JB, He H, Elling AA, Deng XW (2009) Dynamic landscapes of four histone modifications during de etiolation in *Arabidopsis*. *Plant Cell* 21:3732–3748
15. Offermann S, Danker T, Dreytmuller D, Kalamajka R, Topsch S, Weyand K, Peterhansel C (2006) Illumination is necessary and sufficient to induce histone acetylation independent of transcriptional activity at the C4-specific phosphoenolpyruvate carboxylase promoter in maize. *Plant Physiol* 141:1078–1088
16. Ding Y, Avramova Z, Fromm M (2011) The *Arabidopsis* trithorax-like factor ATX1 functions in dehydration stress responses via ABA-dependent and ABA-independent pathways. *Plant J* 66:735–744
17. Song Y, Ji D, Li S, Wang P, Li Q, Xiang F (2012) The dynamic changes of DNA methylation and histone modifications of salt responsive transcription factor genes in soybean. *PLoS One* 7:e41274
18. Kwon CS, Lee D, Choi G, Chung WI (2009) Histone occupancy-dependent and - independent removal of H3K27 trimethylation at cold-responsive genes in *Arabidopsis*. *Plant J* 60:112–121
19. Folsom JJ, Begcy K, Hao X, Wang D, Walia H (2014) Rice fertilization-independent endosperm1 regulates seed size under heat stress by controlling early endosperm development. *Plant Physiol* 165:238–248
20. Köhler C, Makarevich G (2006) Epigenetic mechanisms governing seed development in plants. *EMBO Rep* 7:1223–1227
21. Rodrigues JCM, Johnson S, Okada T, Koltunow AMG (2006) In: Proc. of 8th international congress of plant molecular biology (Apomixis Workshop), Adelaide, Australia
22. Wassenegger M, Heimes S, Riedel L, Sanger HL (1994) RNA-directed de novo methylation of genomic sequences in plants. *Cell* 76:567–576
23. Pikaard CS, Haag JR, Pontes OM, Blevins T, Cocklin R (2012) A transcription fork model for Pol IV and Pol V-dependent RNA-directed DNA methylation. *Cold Spring Harb Symp Quant Biol* 77:205–212
24. Gao ZH et al (2010) An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. *Nature* 465:106–109
25. Buermans HPI, den Dunnen JT (2014) Next generation sequencing technology: advance and applications. *Biochem Biophys Acta* 1842:1932–1941
26. Eichten SR, Springer NM (2015) Minimal evidence for consistent changes in maize DNA methylation patterns following environmental stress. *Front Plant Sci* 6(308)
27. Bilichak A, Ilnytsky Y, Hollunder J, Kovalchuk I (2012) The progeny of *Arabidopsis thaliana* plants exposed to salt exhibit changes in DNA methylation, histone modifications and gene expression. *PLoS One* 7:e30515
28. Li Q et al (2014) Genetic perturbation of the maize methylome. *Plant Cell* 26:4602–4616
29. Zhao L et al (2014) Transcriptional regulation of cell cycle genes in response to abiotic stresses correlates with dynamic changes in histone modifications in maize. *PLoS One* 9:e106070

30. Roy D, Paul A, Roy A, Ghosh R, Ganguly P, Chaudhuri S (2014) Differential acetylation of histone H3 at the regulatory region of OsDREB1b promoter facilitates chromatin remodelling and transcription activation during cold stress. *PLoS One* 9:e100343
31. Hu Y, Zhang L, Zhao L, Li J, He S, Zhou K, Yang F, Huang M, Jiang L, Li L (2011) Trichostatin, A selectively suppresses the cold-induced transcription of the ZmDREB1 gene in maize. *PLoS One* 6:e22132
32. Migicovsky Z, Yao Y, Kovalchuk I (2014) Transgenerational phenotypic and epigenetic changes in response to heat stress in *Arabidopsis thaliana*. *Plant Signal Behav* 9:e27971
33. Sani E, Herzyk P, Perrella G, Colot V, Amtmann A (2013) Hyperosmotic priming of *Arabidopsis* seedlings establishes a long-term somatic memory accompanied by specific changes of the epigenome. *Genome Biol* 14:R59
34. Widiez T, Symeonidi A, Luo C, Lam E, Lawton M, Rensing SA (2014) The chromatin landscape of the moss *Physcomitrella patens* and its dynamics during development and drought stress. *Plant J* 79:67–81
35. Zong W, Zhong X, You J, Xiong L (2013) Genome-wide profiling of histone H3K4-trimethylation and gene expression in rice under drought stress. *Plant Mol Biol* 81:175–188
36. Schneeberger K et al (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc Natl Acad Sci U S A* 108:10249–10254
37. González RM, Ricardi MM, Iusem ND (2013) Epigenetic marks in an adaptive water stress-responsive gene in tomato roots under normal and drought conditions. *Epigenetics* 8:864–872
38. Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X et al (2013) CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res* 23:628–637
39. Stroud H, Ding B, Simon SA, Feng S, Bellizzi M, Pellegrini M, Wang G-L, Meyers BC, Jacobsen SE (2013) Plants regenerated from tissue culture contain stable epigenome changes in rice. *eLife* 2:e00354
40. Schmitz RJ, He Y, Valdés-López O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G, Ecker JR (2013) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res* 23:1663–1674
41. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci U S A* 110:1797–1802
42. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D (2014) Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet* 10:e1004785
43. Karan R, DeLeon T, Biradar H, Subudhi PK (2012) Salt stress induced variation in DNA methylation pattern and its influence on gene expression in contrasting rice genotypes. *PLoS One* 7:e40203
44. Wang W, Zhao X, Pan Y, Zhu L, Fu B, Li Z (2011) DNA methylation changes detected by methylation-sensitive amplified polymorphism in two contrasting rice genotypes under salt stress. *J Genet Genomics* 38:419–424
45. Wang WS, Pan YJ, Zhao XQ, Dwivedi D, Zhu LH, Ali J, Fu BY, Li ZK (2011) Drought-induced site-specific DNA methylation and its association with drought tolerance in rice (*Oryza sativa* L.) *J Exp Bot* 62:1951–1960
46. Zhong L, Xu Y, Wang J (2009) DNA-methylation changes induced by salt stress in wheat *Triticum aestivum*. *Afr J Biotechnol* 8:6201–6207
47. Tan MP (2010) Analysis of DNA methylation of maize in response to osmotic and salt stress based on methylation-sensitive amplified polymorphism. *Plant Physiol Biochem* 48:21–26
48. Marconi G, Pace R, Traini A, Raggi L, Lutts S, Chiusano M, Guiducci M, Falcinelli M, Benincasa P, Albertini E (2013) Use of MSAP markers to analyse the effects of salt stress on DNA methylation in rapeseed (*Brassica napus* var. *oleifera*). *PLoS One* 8:e75597
49. Zheng X, Chen L, Li M, Lou Q, Xia H, Wang P, Li T, Liu H, Luo L (2013) Transgenerational variations in DNA methylation induced by drought stress in two rice varieties with distinguished difference to drought resistance. *PLoS One* 8:e80253

50. Gayacharan A, Joel AJ (2013) Epigenetic responses to drought stress in rice (*Oryza sativa* L.) *Physiol Mol Biol Plants* 19:379–387
51. Gao G, Li J, Li H, Li F, Xu K, Yan G, Chen B, Qiao J, Wu X (2014) Comparison of the heat stress induced variations in DNA methylation between heat-tolerant and heat-sensitive rape-seed seedlings. *Breed Sci* 64:125–133
52. Baranek M, Cechova J, Raddova J, Holleinova V, Ondrusikova E, Pidra M (2015) Dynamics and reversibility of the DNA methylation landscape of grapevine plants (*Vitis vinifera*) stressed by in vitro cultivation and thermotherapy. *PLoS One* 10:e0126638
53. Ou X, Zhang Y, Xu C (2012) Transgenerational inheritance of modified DNA methylation patterns and enhanced tolerance induced by heavy metal stress in rice (*Oryza sativa* L.) *PLoS One* 7:e41143
54. Cicatelli A, Todeschini V, Lingua G, Biondi S, Torrigiani P, Castiglione S (2014) Epigenetic control of heavy metal stress response in mycorrhizal versus non-mycorrhizal poplar plants. *Environ Sci Pollut Res Int* 21:1723–1737
55. Greco M, Chiappetta A, Bruno L, Bitonti MB (2012) In Posidoniaoceanica cadmium induces changes in DNA methylation and chromatin patterning. *J Exp Bot* 63:695–709
56. Yang JL, Liu LW, Gong YQ, Huang DQ, Wang F, He LL (2007) Analysis of genomic DNA methylation level in radish under cadmium stress by methylation-sensitive amplified polymorphism technique. *J Exp Bot* 33:219–226
57. Filek M, Keskinen R, Hartikainen H, Szarejko I, Janiak A, Miszalski Z, Golda A (2008) The protective role of selenium in rape seedlings subjected to cadmium stress. *J Plant Physiol* 165:833–844
58. Rakei A, Maali-Amiri R, Zeinali H, Ranjbar M (2015) DNA methylation and physio-biochemical analysis of chickpea in response to cold stress. *Protoplasma* 253:61–76
59. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD et al (2014) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452:215–219
60. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133:523–536
61. Eichten SR et al (2011) Heritable epigenetic variation among maize inbreds. *PLoS Genet* 7:e1002372
62. Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, Hirsch CN, Dawe RK, Madzima TF, McGinnis KM, Lisch D, Schmitz RJ, Vaughn MW, Springer NM (2015) RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc Natl Acad Sci U S A* 112:14728–14733
63. Takuno S, Ran JH, Gaut BS (2016) Evolutionary patterns of genic DNA methylation vary across land plants. *Nat Plants* 2:15222
64. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328:916–919
65. Zhong S, Fei Z, Chen YR, Zheng Y, Huang M, Vrebalov J, McQuinn R, Gapper N, Liu B, Xiang J, Shao Y, Giovannoni JJ (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat Biotechnol* 31:154–159
66. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A et al (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454:766–770
67. Chen X, Ge X, Wang J, Tan C, King GJ, Liu K (2015) Genome-wide DNA methylation profiling by modified reduced representation bisulfite sequencing in Brassica rapa suggests that epigenetic modifications play a key role in polyploid genome evolution. *Front Plant Sci* 6:836
68. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 13:R87

69. Jiang P, Sun K, Lun FMF, Guo AM, Wang H, Chan KCA et al (2014) Methy-pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. *PLoS One* 9:e100360
70. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27:1571–1572
71. Chen PY, Cokus SJ, Pellegrini M (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* 11:203–208
72. Lim JQ, Tennakoon C, Li G, Wong E, Ruan Y et al (2012) BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. *Genome Biol* 13:R82
73. Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPPING program. *BMC Bioinformatics* 10:232
74. Benoukraf T, Wongphayak S, Hadi LH, Wu M, Soong R (2013) GBSA: a comprehensive software for analysing whole genome bisulfite sequencing data. *Nucleic Acids Res* 41:e55
75. Hansen KD, Langmead B, Irizarry RA (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 13:R83
76. Zhang R et al (2014) PEAR: a fast and accurate Illumina Paired-End reAd merge. *Bioinformatics* 30:614–620
77. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord RV, Clark SJ, Molloy PL (2015) De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* 8:6
78. Song Q, Decato B, Hong E, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* 8:e81148
79. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30:1363–1369
80. Akman K (2014) BEAT: BEAT - BS-Seq epimutation analysis toolkit. R package version 1.8.0
81. Martin TC, Yet I, Tsai PC, Bell JT (2015) coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinformatics* 16:131
82. Statham AL, Strbenac D, Coolen MW, Stirzaker C, Clark SJ, Robinson MD (2010) Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics* 26:1659
83. Kishore K, de Pretis S, Lister R, Morelli MJ, Bianchi V, Amati B, Ecker JR, Pelizzola M (2015) methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. *BMC Bioinformatics* 16:313–324
84. Klein H, Schaefer M, Porse BT, Hasemann MS, Ickstadt K, Dugas M (2014) Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. *Bioinformatics* 30:1154–1162
85. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model based analysis of ChIP-Seq (MACS). *Genome Biol* 9:R137
86. Schweikert G (2012) MMDiff: statistical testing for ChIP-Seq data sets. R package version 1.10.0
87. Choi CS, Sano H (2007) Abiotic-stress induces demethylation and transcriptional activation of a gene encoding a glycerophosphodiesterase-like protein in tobacco plants. *Mol Genet Genomics* 277:589–600
88. Ocaña J, Walter B, Schellenbaum P (2013) Stable MSAP markers for the distinction of *Vitis vinifera* cv Pinot Noir Clones. *Mol Biotechnol* 55:236–248
89. Danker T (2008) Developmental information but not promoter activity controls the methylation state of histone H3 lysine 4 on two photosynthetic genes in maize. *Plant J* 53:465–474
90. Jaskiewicz M (2011) Chromatin modification acts as a memory for systemic acquired resistance in the plant stress response. *EMBO Rep* 12:50–55

91. Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol* 10:R62
92. van Dijk K, Ding Y, Malkaram S, Riethoven JJM, Liu R, Yang J, Laczko P, Chen H, Xia Y, Ladunga I et al (2010) Dynamic changes in genome-wide histone H3 lysine 4 methylation patterns in response to dehydration stress in *Arabidopsis thaliana*. *BMC Plant Biol* 10:238
93. Mosquna A, Katz A, Decker EL, Rensing SA, Reski R, Ohad N (2009) Regulation of stem cell maintenance by the Polycomb protein FIE has been conserved during land plant evolution. *Development* 136:2433–2444
94. Okano Y, Aono N, Hiwatashi Y, Murata T, Nishiyama T, Ishikawa T, Kubo M, Hasebe M (2009) A Polycomb repressive complex 2 gene regulates apogamy and gives evolutionary insights into early land plant evolution. *Proc Natl Acad Sci U S A* 106:16321–16326
95. Schramke V, Allshire R (2003) Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science* 301:1069–1074
96. Durán-Figueroa N, Vielle-Calzada JP (2010) ARGONAUTE9-dependent silencing of transposable elements in pericentromeric regions of *Arabidopsis*. *Plant Signal Behav* 6:5
97. Schmitz RJ, Zhang X (2011) High-throughput approaches for studying plant epigenomics. *Curr Opin Plant Biol* 14:130–136
98. Boyko A, Kovalchuk I (2008) Epigenetic control of plant stress response. *Environ Mol Mutagen* 49:61–72
99. Chinnusamy V, Zhu JK (2009) Epigenetic regulation of stress responses in plants. *Curr Opin Plant Biol* 12:1–7
100. Zhu J, Jeong JC, Zhu Y, Sokolchik I, Miyazaki S, Zhu JK, Hasegawa PM, Bohnert HJ, Shi H, Yun DJ et al (2008) Involvement of *Arabidopsis* HOS15 in histone deacetylation and cold tolerance. *Proc Natl Acad Sci U S A* 105:4945–4950
101. Sokol A, Kwiatkowska A, Jerzmanowski A, Prymakowska-Bosak M (2007) Up-regulation of stress-inducible genes in tobacco and *Arabidopsis* cells in response to abiotic stresses and ABA treatment correlates with dynamic changes in histone H3 and H4 modifications. *Planta* 227:245–254

Nanotechnology in Plants



Ismail Ocsoy, Didar Tasdemir, Sumeyye Mazicioglu, and Weihong Tan

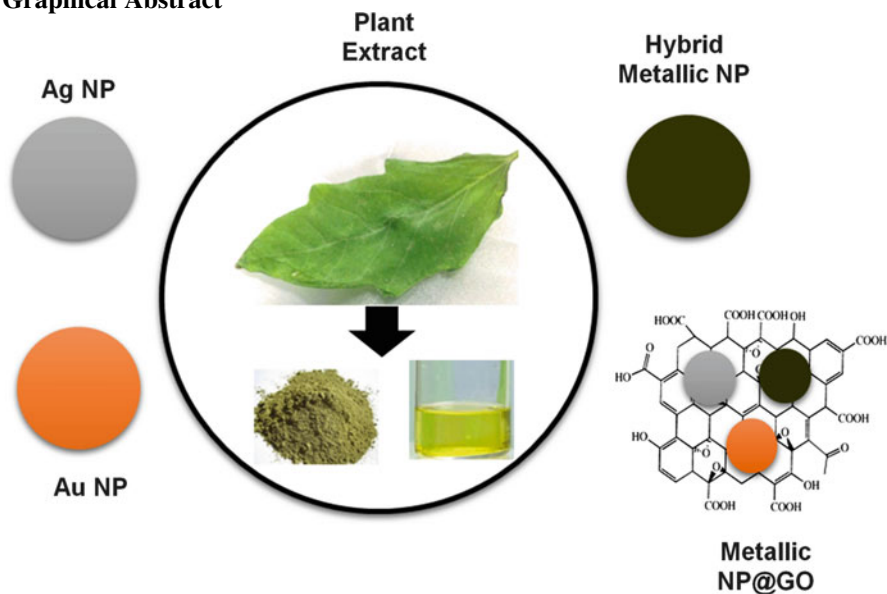
Abstract The integration of nanotechnology in medicine has had a tremendous impact in the past few decades. The discovery of synthesis of nanomaterials (NMs) and their functions as versatile tools promoted various applications in nanobiotechnology and nanomedicine. Although the physical and chemical methods are still considered as commonly used methods, they introduce several drawbacks such as the use of toxic chemicals (solvent, reducing, and capping agents) and poor control of size, size distribution, and morphology, respectively. Additionally, the NMs synthesized in organic solvents and hydrophobic surfactants rapidly aggregate in aqueous solutions or under physiologic conditions, limiting their applications in medicine. Many of the phase-transfer strategies were developed and applied for the transfer of NMs into aqueous solutions. Although great efforts have been put into phase transfers, they mostly include expensive, time-consuming, intensive labor work, multi steps, and complicated procedures.

Use of plant extracts in the biological synthesis method offers stark advantages over other biomolecules (protein, enzyme, peptide, and DNA). Plant extracts have been commonly used for food, medicine, NM synthesis, and biosensing. There are many viable techniques developed for the production of plant extracts with various contents based on their simplicity, cost, and the type of extract content. In this chapter, we conduct a comparative study for extract preparation techniques, the use of extracts for metallic single and hybrid nanoparticle (NP) synthesis, and their antimicrobial properties against pathogenic and plant-based bacteria.

I. Ocsoy (✉), D. Tasdemir, and S. Mazicioglu
Department of Analytical Chemistry, Faculty of Pharmacy, Erciyes University, Kayseri, Turkey
e-mail: ismailocsoy66@gmail.com

W. Tan (✉)
Department of Chemistry and Shands Cancer Center, University of Florida, Gainesville, FL,
USA
e-mail: tan@chem.ufl.edu

Graphical Abstract



Keywords Disease and antimicrobial property, Metallic NP, Microorganisms, Plant extract

Contents

1	Introduction	264
2	A Comparative Study of Extraction Methods for Medicinal Plants	265
3	Preparation of Extract for Nanoparticle (NP) Synthesis	266
3.1	Solvent-based Extraction Methods	266
3.2	Microwave-assisted Extraction Technique	266
4	Isolation of Specific Molecules from Plant Extracts Using Appropriate NPs	267
5	Synthesis of NPs Using Plant Extracts	267
5.1	Synthesis of Silver and Gold NPs	268
6	Plant Disease Treatment	269
6.1	Antimicrobial Properties of Silver Nanoparticles	270
7	Conclusion	270
	References	271

1 Introduction

In recent years, nanotechnology has been increasingly utilized for the synthesis, engineering, and designing of various nanomaterials (NMs) used as antioxidants, antimicrobials, anticancer agents, therapeutics, and diagnostics agents, and in the fabrication of nanosensors. The NMs have been intensively applied in many

different scientific and industrial fields. However, production of biogenic NMs in nanotechnology and their uses in medicine have become the fastest developing and most attractive research. The use of plant extracts is prominent, not only because of their easy production and cheap cost compared to other biomolecules (protein, enzyme, peptide, and DNA), but also because they provide large-scale and environmentally benign NMs, which can widen their medical applications.

2 A Comparative Study of Extraction Methods for Medicinal Plants

Medicinal plants have been practically used as effective and traditional drugs or biocides for various disorders for a long time. The research has focused on the elucidation of the chemical structures of these plant extracts. The physicochemical properties of plants are investigated at the following steps: (1) authentication, (2) extraction, (3) separation, (4) isolation, (5) characterization of isolated compounds, and (6) quantitative evaluation. The methods vary in simplicity, cost, efficiency, and degree of extracted or isolated molecule damage. It is worth mentioning that each extract needs its own characteristic extraction method for production with greater efficiency. For instance, while essential oils as volatile compounds in aromatic plants are extracted using distillation methods, solvent extraction methods are viable and suitable for obtaining other volatile compound-rich extracts.

In addition to conventional extraction methods, some modern extraction methods including microwave-assisted extraction (MAE), ultrasonication-assisted extraction (UAE), supercritical fluid extraction (SFE), and solid-phase micro-extraction (SPME) have been developed and are actively used. They display certain advantages over conventional methods. Although classical methods are fairly simple, standard, and have widespread use, they consume large quantities of organic solvents, cause degradation of heat-labile constituents, produce extracts with a low yield, and have time-consuming and labor-intensive procedures. The use of these classical extraction methods allows the benefits of production efficiency, selectivity, and the elimination of additional steps of modern extraction methods before chromatographic analysis allows them to be used intensively and preferentially. The extraction procedures can also be redesigned to obtain the desired molecules by manipulating experimental parameters. Additionally, the extraction method selection to isolate targeted components with the highest yield and highest purity is dependent upon the plant source [1, 2]. Thus, the development of modern extraction methods plays an important role in the overall effort of ensuring and providing high-quality herbal products [3].

3 Preparation of Extract for Nanoparticle (NP) Synthesis

The extraction of plant contents has received considerable attention owing to the use of plant contents in medicine, nanoparticle (NP) synthesis, and biosensing. Three major extraction methods for NP synthesis are: (1) solvent extraction, (2) microwave-assisted extraction, and (3) maceration extraction. The ideal extraction method should be cost-effective, simple, less time-consuming, and simply conducted in any laboratory.

3.1 Solvent-based Extraction Methods

The solid-liquid extraction provides soluble components in the solid material to be integrated with the solvent. The mass transfer ratio decreases as the concentration of the active principle in the solvent increases. This process results in the solvent and solid material reaching an equilibrium concentration when a mass transfer of the active components from plant material to solvent occurs. There are different types of this technique: cold percolation, hot percolation, and concentration [4, 5].

3.1.1 Cold Percolation

The extraction of the plant contents is carried out in a percolator that is connected to a condenser and a receiver for removing the solvent from the mixture. The powdered material is in contact with the percolator along with a suitable solvent until equilibrium is reached.

3.1.2 Hot Percolation

The principle of hot percolation is based on increasing the temperature of the solvent, which increases the solubility. The extract is permanently passed into a tubular heat exchanger by steam heating.

3.2 Microwave-assisted Extraction Technique

Microwave-assisted extraction (MAE) allows the materials to reach the given energy that is associated with the dielectric susceptibility of both the solvent and the solid plant material, through rapid heating [6, 7].

3.2.1 Maceration

The maceration process can be completed in three steps: (1) grinding of plant materials into small particles, (2) choosing the appropriate solvent, which is added in a closed vessel, and (3) filtration for the separation of the liquid phase from the plant pulp. The mixture of the plant powder and solvent should be shaken to provide proper extraction with a high efficiency [8, 9].

4 Isolation of Specific Molecules from Plant Extracts Using Appropriate NPs

It is well known that titanium dioxide (TiO_2), with a certain crystal form, potentially reacts with phosphorylated biomolecules including peptides, proteins, and glycoproteins, under the proper acidic experimental conditions. The phosphate moieties of those molecules specifically attach to the surface of TiO_2 NPs and retain their surfaces, which provides the isolation or enrichment of the corresponding molecules.

Recently, TiO_2 NPs have been integrated into plant nanobiology for isolation of specific molecules from plants. In a recent studies, TiO_2 NPs acted as nano-harvesting agents to isolate bioactive compounds from living cells. For instance, Kurepa et al. [10] used phosphorylated anatase TiO_2 NPs with a 20-nm diameter to capture the specific flavonoids from Arabidopsis plants. This work showed that quercetin and kaempferol as enediol and catechol groups containing flavonoids can successfully bind to the phosphorylated TiO_2 NPs, and they were isolated from the plant matrix.

5 Synthesis of NPs Using Plant Extracts

The production of colloidal metallic NPs has become one of the fastest developing and most exciting fields of research and has had an enormous impact on the evolution of nanotechnology over the past decades. The size, shape, and composition-dependent electronic, optical, luminescent, and magnetic features of the NPs with great enhancement have found a wide spectrum of applications in scientific and technical fields [11–16]. In general, three major methods, chemical, physical, and biological, have been actively and extensively used for the synthesis of NPs. Although chemical methods are used the most for the synthesis of high-quality NPs with a narrow size distribution, the use of toxic organic solvents as well as reducing and stabilizing agents greatly limits the applications for NPs, especially in biomedicine and bioanalytics [16–19]. Additionally, in order to use NPs synthesized in organic solvents in biologically-related applications, phase transfer is an indispensable step for introducing the NPs into aqueous solutions. The NPs can be made

water-soluble with two common surface-engineering procedures: (1) ligand exchange and (2) ligand polymerization [20–25].

Physical methods include simple one-step procedures and provide large-scale production in a short time. However, those methods almost always result in a lack of size, shape, and size distribution of the NPs [26, 27]. To address the drawbacks encountered in chemical and physical methods, researchers have recently focused on biological methods called “green methods” for NP synthesis [28, 29].

The main principle of green methods is to use nontoxic biomolecules including DNA, proteins, enzymes, carbohydrates, and plant extracts for the synthesis of biocompatible metallic NPs through the reduction of metal ions in aqueous solution [30–38]. Although DNA, proteins, and enzymes have been employed as scaffolds for nucleation and growth of metallic NPs with unique crystalline structures [39–43], those biomolecules are quite costly, easily decomposable, and can be contaminated. In contrast to those molecules, plant extracts are easily reachable, quite affordable, and very stable against environmental conditions (temperature, pH, and salt concentration).

Plant extracts are a rich source of polyphenols, flavonoids, sugars, enzymes, and/or proteins, and can be utilized as reducing and stabilizing agents for the biosynthesis of metallic NPs. In the potential proposed mechanism, hydroxyl, amine, and thiols groups existing on particular extracts of plants may bind to metal ions, canalize electron flow from the extracts to metal ions, and lead to the completion of the eventual NP synthesis [44–49]. Numerous numbers of plant extract-directed metallic NPs have been synthesized and used in various fields [47–49]. Using plant extracts for the rapid reduction and formation of metallic NPs was discovered by Sastry and co-workers. They used lemongrass plant extract to synthesize the spherical gold NPs and triangular gold nanoprisms [50]. In addition to the aforementioned unique properties of plant extracts, plant-based biogenic synthesis can provide cost-effective, environmentally-friendly, simple, less labor intensive, and large-scale production procedures.

5.1 Synthesis of Silver and Gold NPs

Silver (Ag) and gold (Au) are two commonly synthesized plasmonic NPs, due to their unique intrinsic properties. Ag NPs have been considered as effective and universal germicidal agents against various microbes. Their use has also been recognized in nanomedical and industrial applications [51–54]. As stated above, the plant extract-based synthesis method for Ag NPs provides simple, one-step, and rapid procedures compared to other synthesis methods. The extracts produced from different parts of the plant such as leaves, roots, seeds, and fruit act as potential reducing and stabilizing agents to form Ag NPs of various sizes and shapes. For instance, square, spherical, triangular, and hexagonal-shaped Ag NPs with diameters ranging from 10 to 90 nm were synthesized using leaf extracts obtained from plants [55–57]. In our view, different plants contain different contents in the extract, which

may lead to the formation of Ag NPs of various sizes and shapes. The synthesis of Ag NPs of various morphologies was systematically reported using extracts of roots, seeds, and fruit [58–60].

Similar strategies have been used for the synthesis of Au NPs. Au NPs have received considerable attention due to their attractive optical and non-toxic properties, which market them for use in a wide variety of scientific areas including nanomedicine, nanoelectronics, nanobiosensing, and catalysis [61–66]. Similar to Ag NPs, the extracts obtained from different parts of plants reacted with Au ions to reduce them and to eventually form Au NPs. For instance, while leaf extract from the *Menta piperita* plant resulted in the formation of spherical Au NPs with a diameter of 150 nm, triangular-, hexagonal-, and pentagonal-shaped Au NPs with a size ranging from 5 to 500 nm were synthesized from the extracts of *Coriandum sativum*, *Memecylon edule*, and *Magnolia kobus* plants [55, 67, 68].

6 Plant Disease Treatment

Various techniques have been developed and applied to control microorganism-caused diseases in plants. For instance, the antibiotic streptomycin, as part of a chemical technique, was used in the 1950s to prevent the proliferation of *Xanthomonas vesicatoria* found on plants. However, those bacterial strains developed resistance against streptomycin and thus made it ineffective [69]. Copper-based (Cu) bactericides incorporated ethylene-bis-dithiocarbamate (EBDC) fungicides (e.g., maneb or mancozeb). These fungicides (e.g., maneb or mancozeb) have acted as potential biocides in order to effectively manage the diseases existing on plants. Nevertheless, Cu-resistant bacterial strains have been observed due to their frequent use and the resulting drastic reduction in antimicrobial activity of those biocides [70–72].

Research has focused on investigation and development of bacteriophages and systemic-acquired resistance (SAR) inducers as alternative disease-management techniques over the last decade [73, 74]. As an example, acibenzolar-S-methyl (ASM) was used as an SAR inducer agent, to activate and enhance plant defense systems by increasing the transcription of stress-related genes against bacterial tomato spot [73]. Although bacteriophages have been introduced as biological alternatives to Cu-based bactericides, real-time use in the field reduced their viability and then their use was highly limited due to environmental conditions [75, 76]. It is worth mentioning that only very few chemical techniques are available and there is an urgent need to develop effective, biocompatible, and economical materials for disease management.

No reports have fully explained the mechanism underlying the antimicrobial activity of NPs, and the mechanism is still under debate. Recently, various types of single-component metallic NPs and metal-graphene oxide (GO) NPs have been synthesized and used as novel and effective antimicrobial agents for the management of agricultural crop diseases. The key point in the use of NPs is their toxicity, which

can adversely influence environmental and human health [77–80]). For instance, Paret et al. studied antibacterial properties of light-activated titanium dioxide (TiO_2) and metal-doped hybrid TiO_2 NPs (TiO_2/Ag , TiO_2/Zn) against *Xanthomonas perforans*, which causes bacterial tomato spot disease. This study demonstrated that TiO_2 did not show any antimicrobial function under non-illuminated conditions and only TiO_2/Ag exhibited some antimicrobial activity due to the intrinsic antimicrobial property of Ag. In contrast, all TiO_2 -based NPs effectively inhibited bacterial growth when exposed to an incandescent light intensity of 3×10^4 lux. The combination of the photocatalytic activity of TiO_2 and the natural germicidal activity of Ag introduced the best antimicrobial activity under illuminated conditions [80].

6.1 Antimicrobial Properties of Silver Nanoparticles

Silver NPs have been considered to be the strongest and most universal biocides compared to other metallic NPs. The one logical proposed mechanism offered is that Ag NP may interact with some functional groups (thiol, carboxyl, hydroxyl, amino, and phosphate groups) existing on bacterial membranes, with membrane degradation then leading to serious structural deformation. In addition to that, some Ag NPs can be internalized through the membranes and may inactivate or distort the working function of enzymes, which may lead to cell death [81, 82]. However, when Ag NPs are aggregated, their antimicrobial activities are weakened and can be lost. Most recent works show that Ag-GO nanocomposites overcome the limitations of bare Ag NPs. Ag-GO nanocomposites display extraordinary antibacterial activity that results in rapid killing [83, 84].

7 Conclusion

The type of extraction method used varies according to the type of content in the extracts. The use of plant extracts has advantages over other biomolecules (proteins, enzymes, peptides, and DNA) in terms of the biosynthesis of metallic NPs, because they are inexpensive, easily producible, and accessible. They provide environmentally friendly NPs with the ability for large-scale production. For these reasons, plant extract-directed NPs can potentially be used in various bioanalytical and biomedical applications as antioxidants, antimicrobial agents, anticancer agents, therapeutics, diagnostic tools, and drug-vehicle agents.

References

1. Jain D, Daima HK, Kachhwaha S, Kothari SL (2009) Synthesis of plant-mediated silver nanoparticles using papaya fruit extract and evaluation of their anti microbial activities. *Dig J Nanomater Biostruct* 4:557–563
2. Gupta A, Naranawal M, Kothari V (2012) Modern extraction methods for preparation of bioactive plant extracts. *IJANS* 1:8–26
3. Huie CW (2002) A review of modern sample-preparation techniques for the extraction and analysis of medicinal plants. *Anal Bioanal Chem* 373:23–30
4. Chen S, Sun Y, Chao J, Cheng L, Chen Y, Liu J (2011) Dispersive liquid–liquid microextraction of silver nanoparticles in water using ionic liquid 1-octyl-3 methylimidazolium hexafluorophosphate. *J Environ Sci* 41:211–217
5. Nerome H, Machmudah S, Fukuzato R, Higashiura T, Kanda H, Goto M (2016) Effect of solvent on nanoparticle production of β -carotene by a supercritical anti-solvent process. *Chem Eng Technol* 39:1771–1777
6. Surendra TV, Roopan SM, Arasu MV, Al-Dhabi NA, Rayalu GM (2016) RSM optimized Moringa oleifera peel extract for green synthesis of M. oleifera capped palladium nanoparticles with antibacterial and hemolytic property. *J Photochem Photobiol B* 162:550–557
7. Sharma D, Sabela MI, Kanchi S, Mdluli PS, Singh G, Stenström TA, Bisetty K (2016) Biosynthesis of ZnO nanoparticles using Jacaranda mimosifolia flowers extract: synergistic antibacterial activity and molecular simulated facet specific adsorption studies. *J Photochem Photobiol B* 162:199–207
8. Chandran SP, Chaudhary M, Pasricha R, Ahmad A, Sastry M (2006) Synthesis of gold nano-triangles and silver nanoparticles using Aloe vera plant extract. *Biotechnol Prog* 22:577–583
9. Azmir J, Zaidul ISM, Rahman MM, Sharif KM, Mohamed A, Sahena F, Omar AKM (2013) Techniques for extraction of bioactive compounds from plant materials: a review. *J Food Eng* 117:426–436
10. Kurepa J, Nakabayashi R, Paunesku T, Suzuki M, Saito K, Woloschak GE, Smalle JA (2014) Direct isolation of flavonoids from plants using ultra-small anatase TiO₂ nanoparticles. *Plant J* 77:443–453
11. Ma X, Zhao Y, Liang X-J (2011) Theranostic nanoparticles engineered for clinic and pharmaceuticals. *Acc Chem Res* 44:1114–1122
12. Wang H, Yang R, Yang L, Tan W (2009) Nucleic acid conjugated nanomaterials for enhanced molecular recognition. *ACS Nano* 3:2451–2460
13. Hu R, Zhang X-B, Kong R-M, Zhao X-H, Jiang J, Tan W (2011) Nucleic acid-functionalized nanomaterials for bioimaging applications. *J Mater Chem* 21:16323–16334
14. Shrivastava K, Wu H-F (2010) Multifunctional nanoparticles composite for MALDI-MS: Cd₂s-doped carbon nanotubes with CdS nanoparticles as the matrix, preconcentrating and accelerating probes of microwave enzymatic digestion of peptides and proteins for direct MALDI-MS analysis. *J Mass Spectrom* 45:1452–1460
15. Murray C-B, Norris D-J, Bawendi M-G (1993) Synthesis and characterization of nearly monodisperse CdE (E = Sulfur, selenium, tellurium) semiconductor nanocrystallites. *J Am Chem Soc* 115:8706–8715
16. Rosenthal S-J, Chang J-C, Kovtun O, McBride J-R, Tomlinson I-D (2011) Biocompatible quantum dots for biological applications. *Chem Biol* 18:10–24
17. Sun S, Zeng H, Robinson DB, Raoux S, Rice PM, Wang SX, Li G (2004) Monodisperse MFe₂O₄ (M = Fe, Co, Mn) nanoparticles. *J Am Chem Soc* 126:273–279
18. Villaraza A-J, Bump A, Brechbiel M-W (2010) Macromolecules, dendrimers, and nanomaterials in magnetic resonance imaging: the interplay between size, function, and pharmacokinetics. *Chem Rev* 110:2921–2959
19. Agnihotri S, Mukherji S, Mukherji S (2014) Size-controlled silver nanoparticles synthesized over the range 5–100 nm using the same protocol and their antibacterial efficacy. *RSC Adv* 4:3974

20. Michalet X, Pinaud F-F, Bentolila L-A, Tsay J-M, Doose S, Li J-J, Sundaresan G, Wu A-M, Gambhir S-S, Weiss S (2005) Quantum dots for live cells, in vivo imaging, and diagnostics. *Science* 307:538–544
21. Sperling R-A, Parak W-J (2010) Surface modification, functionalization and bioconjugation of colloidal inorganic nanoparticles. *Philos Trans R Soc Lond Ser A* 368:1333–1383
22. Lin C-AJ, Sperling R-A, Li J-K, Yang T-Y, Li P-Y, Zanella M, Chang W-H, Parak W-J (2008) Design of an amphiphilic polymer for nanoparticle coating and functionalization. *Small* 4: 334–341
23. Chen T, Ocsoy I, Yuan Q, Wang R, You M, Zhao Z, Song E, Zhang X, Tan W (2012) One-step facile surface engineering of hydrophobic nanocrystals with designer molecular recognition. *J Am Chem Soc* 134:13164–13167
24. Ocsoy I, Gulbakan B, Shukoor M-I, Xiong X, Chen T, Powell D-H, Tan W (2013) Aptamer-conjugated multifunctional Nnanoflowers as a platform for targeting, apture, and detection in laser desorption ionization mass spectrometry. *ACS Nano* 7:417–427
25. Peng L, You M, Wu C, Han D, Öcsoy I, Chen T, Chen Z, Tan W (2014) Reversible phase transfer of nanoparticles based on photoswitchable host–guest chemistry. *ACS Nano* 8: 2555–2561
26. Herzer G (1989) Grain structure and magnetism of nanocrystalline ferromagnets. *IEEE Trans Magn* 25:3327–3329
27. Skorvánek I, O’Handley R-C (1995) Fine-particle magnetism in nanocrystalline Fe-CuNb-Si-B at elevated temperatures. *J Magn Magn Mater* 140–144:467–468
28. Raveendran P, Fu J, Wallen S-L (2003) Completely “green” synthesis and stabilization of metal nanoparticles. *J Am Chem Soc* 125:13940–13941
29. Irvani S (2011) Green synthesis of metal nanoparticles using plants. *Green Chem* 13:50–2638
30. De La Rica R, Matsui H (2008) Urease as a nanoreactor for growing crystalline ZnO nanoshells at room temperature. *Angew Chem Int Ed* 47:5415–5417
31. Ocsoy I, Gulbakan B, Chen T, Zhu G, Chen Z, Sari M-M, Peng L, Xiong X, Fang X, Tan W (2013) DNA-guided metal-nanoparticle formation on graphene oxide surface. *Adv Mater* 25: 2319–2325
32. Ocsoy I, Paret M-L, Ocsoy M-A, Kunwar S, Chen T, You M, Tan W (2013) Nanotechnology in plant disease management: DNA-directed silver nanoparticles on graphene oxide as an anti-bacterial against *xanthomonas perforans*. *ACS Nano* 7:8972–8980
33. Li C, Chen T, Ocsoy I, Zhu G, Yasun E, You M, Wu C, Zheng J, Song E, Huang C-Z, Tan W (2014) Gold-coated Fe₃O₄ nanoroses with five unique functions for cancer cell targeting, imaging and therapy. *Adv Funct Mater* 24:1772–1780
34. Leng Y, Fu L, Ye L, Li B, Xu X, Xing X, He J, Song Y, Leng C, Guo Y, Ji X, Lu Z (2016) Protein-directed synthesis of highly monodispersed, spherical gold nanoparticles and their applications in multidimensional sensing. *Sci Rep* 6:28900
35. Strayer A-L, Ocsoy I, Tan W, Jones J, Paret M-L (2016) Low concentrations of a silver-based nanocomposite to manage bacterial spot of tomato in the greenhouse. *Plant Dis* 100:1460–1465
36. Duman F, Ocsoy I, Kup F-O (2016) Chamomile flower extract-directed CuO nanoparticle formation for its antioxidant and DNA cleavage properties. *Mat Sci Eng C* 60:333–338
37. Demirbas A, Welt B-A, Ocsoy I (2016) Biosynthesis of red cabbage extract directed Ag NPs and their effect on the loss of antioxidant activity. *Mater Lett* 179:20–23
38. Sun Q, Cai X, Li J, Zheng M, Chen Z, Yu C-P (2014) Green synthesis of silver nanoparticles using tea leaf extract and evaluation of their stability and antibacterial activity. *Colloids Surf A Physicochem Eng Asp* 444:226–231
39. Wei H, Wang Z, Zhang J, House S, Gao Y-G, Yang L, Robinson H, Tan L-H, Xing H, Hou C, Robertson I-M, Zuo J-M, Lu Y (2011) Time-dependent, protein-directed growth of gold nanoparticles within a single crystal of lysozyme. *Nat Nanotechnol* 6:93–97
40. Ma X, Huh J, Park W, Lee L-P, Kwon Y-J, Sim S-J (2016) Gold nanocrystals with DNA-directed morphologies. *Nat Commun* 7:12873

41. Rodríguez-Lorenzo L, De La Rica R, Álvarez-Puebla R-A, Liz-Marzán L-M, Stevens M-M (2012) Plasmonic nanosensors with inverse sensitivity by means of enzyme-guided crystal growth. *Nat Mater* 11:604–607
42. Tikhomirov G, Hoogland S, Lee P-E, Fischer A, Sargent E-H, Kelley S-O (2011) DNA-based programming of quantum dot valency, self-assembly and luminescence. *Nat Nanotechnol* 6:485–490
43. Ma N, Sargent E-H, Kelley S-O (2009) One-step DNA-programmed growth of luminescent and biofunctionalized nanocrystals. *Nat Nanotechnol* 4:121–125
44. Karatoprak G-Ş, Aydin G, Altinsoy B, Altinkaynak C, Koşar M, Ocsoy I (2017) The effect of pelargonium *Endlicherianum* fenzl. Root extracts on formation of nanoparticles and their antimicrobial activities. *Enzyme Microb Technol* 97:21–26
45. Katircioğlu Z, Şakalaka H, Ulaşan M, Gören A-C, Yavuz M-S (2014) Facile synthesis of “green” gold nanocrystals using cynarin in an aqueous solution. *Appl Surf Sci* 318:191–198
46. Ocsoy I, Temiz M, Celik C, Altinsoy B, Yilmaz V, Duman F (2017) A green approach for formation of silver nanoparticles on magnetic graphene oxide and highly effective antimicrobial activity and reusability. *J Mol Liq* 227:147–152
47. Mittal A-K, Chisti Y, Banerjee U-C (2013) Synthesis of metallic nanoparticles using plant extracts. *Biotechnol Adv* 31:346–356
48. Akhtar M-S, Panwar J, Yun Y-S (2013) Biogenic synthesis of metallic nanoparticles by plant extracts. *ACS Sustain Chem Eng* 1:591–602
49. Park Y, Hing Y-N, Weyers A, Kim Y-S, Linhardt R-J (2011) Polysaccharide and phytochemicals: a natural reservoir for the green synthesis of gold and silver nanoparticles. *IET Nano-biotechnol* 5:69–78
50. Shankar S-S, Rai A, Ankamwar B, Singh A, Ahmad A, Sastry M (2004) Biological synthesis of triangular gold nanoprisms. *Nat Mater* 3:482
51. Jiang H, Manolache S, Wong ACL, Denes FS (2004) Plasmaenhanced deposition of silver nanoparticles onto polymer and metal surfaces for the generation of antimicrobial characteristics. *J Appl Polym Sci* 93(3):1411–1422
52. Sondi I, Salopek-Sondi B (2004) Silver nanoparticles as antimicrobial agent: a case study on *E. coli* as a model for gram-negative bacteria. *J Colloid Interface Sci* 275:177–182
53. Kim K-J, Sung W, Suh B, Moon S-K, Choi J-S, Kim J, Lee D (2009) Antifungal activity and mode of action of silver nano-particles on *Candida albicans*. *Biometals* 22:235–242
54. Zodrow K, Brunet L, Mahendra S, Li D, Zhang A, Li Q, Alvarez PJJ (2009) Polysulfone ultrafiltration membranes impregnated with silver nanoparticles show improved biofouling resistance and virus removal. *Water Res* 43:715–723
55. Elavazhagan T, Arunachalam KD (2011) Memecylon edule leaf extract mediated green synthesis of silver and gold nanoparticles. *Int J Nanomedicine* 6:1265–1278
56. Philip D, Unni C, Aromal SA, Vidhu VK (2011) *Murraya koenigii* leaf-assisted rapid green synthesis of silver and gold nanoparticles. *Spectrochim Acta Part A* 78(2):899–904
57. Phillip D (2011) *Mangifera indica* leaf-assisted biosynthesis of welldispersed silver nanoparticles. *Spectrochim Acta Part A* 78(1):327–331
58. Bar H, Bhui DK, Sahoo GP, Sarkar P, Pyne S, Misra A (2009) Green synthesis of silver nanoparticles using seed extract of *Jatropha curcas*. *Colloids Surf A Physicochem Eng Asp* 348: 212–216
59. Ahmad N, Sharma S, Alam MK, Singh VN, Shamsi SF, Mehta BR, Fatma A (2010) Rapid synthesis of silver nanoparticles using dried medicinal plant of basil. *Colloids Surf B Bio-interfaces* 81(1):81–86
60. Dubey SP, Lahtinen M, Sillanpaa M (2010) Tansy fruit mediated greener synthesis of silver and gold nanoparticles. *Process Biochem* 45(7):1065–1071
61. Aromal SA, Philip D (2012) Green synthesis of gold nanoparticles using *Trigonella foenum-graceum* and its size-dependent catalytic activity. *Spectrochim Acta Part A* 97:1–5

62. Liping Q, Tao C, Ismail Ö, Emir Y, Wu C, Guizhi Z, Mingxu Y, Da H, Jianhui J, Ruqin Y, Weihong T (2015) A cell-targeted, size-photocontrollable, nuclear-uptake nanodrug delivery system for drug-resistant cancer therapy. *Nano Lett* 15:457–463
63. Yasun E, Gulbakan B, Ocsoy I, Yuan Q, Shukoor MI, Li C, Tan W (2012) Enrichment and detection of rare proteins with aptamer-conjugated gold nanorods. *Anal Chem* 84:6008–6015
64. Shukoor MI, Altman MO, Han D, Bayrac AT, Ocsoy I, Zhu Z, Tan W (2012) Aptamer-nanoparticle assembly for logic-based detection. *ACS Appl Mater Interfaces* 4:3007–3011
65. Ocsoy I, Arslan Ocsoy M, Yasun E, Tan W (2013) Nucleic acid-functionalized nanomaterials. *Nano Life* 03:1–10
66. McLamore ES, Convertino M, Ocsoy I, Vanegas DC, Taguchi M, Rong Y, Gomes C, Chaturvedi P, Claussen JC (2016) Biomimetic fractal nanometals as a transducer layer in electrochemical biosensing. *Semiconductor-based sensors*. World Scientific Publishing, Singapore, pp 35–67. https://doi.org/10.1142/9789813146730_0002
67. Narayanan KB, Sakthivel N (2008) Coriander leaf mediated biosynthesis of gold nanoparticles. *Mater Lett* 62(30):4588–4590
68. Song JY, Jang HK, Kim BS (2009) Biological synthesis of gold nanoparticles using *Magnolia kokus* and *Diopyros kaki* leaf extracts. *Process Biochem* 44:1133–1138
69. Thayer PL, Stall RE (1962) A survey of *xanthomonas vesicatoria* resistance to streptomycin. *Proc Fla State Hort Soc* 75:163–165
70. Jones JB, Jones JP (1985) The effect of bactericides, tank mixing time and spray schedule on bacterial leaf spot of tomato. *Proc Fla State Hort Soc* 98:244–247
71. Marco GM, Stall RE (1983) Control of bacterial spot of pepper initiated by strains of *xanthomonas campestris* P.v. *vesicatoria* that differ in sensitivity to copper. *Plant Dis* 67: 779–781
72. Jones JB, Woltz SS, Jones JP, Portier KL (1991) Population dynamics *xanthomonas campestris* P.v. *vesicatoria* on tomato leaflets treated with copper bactericides. *Phytopathology* 81:714–719
73. Obradovic A, Jones JB, Momol MT, Olson SM, Jackson LE, Balogh B, Guven K, Iriarte FB (2005) Integration of biological control agents and systemic acquired resistance inducers against bacterial spot on tomato. *Plant Dis* 89:712–716
74. Huang C-H, Vallad GE, Zhang S, Wen A, Balogh B, Figueiredo JFL, Behlau F, Jones JB, Momol MT, Olson SM (2012) Effect of application frequency and reduced rates of acibenzolar-S-methyl on the field efficacy of induced resistance against bacterial spot on tomato. *Plant Dis* 96:221–227
75. Neal A (2008) What can be inferred from bacterium-nanoparticle interactions about the potential consequences of environmental exposure to nanoparticles? *Ecotoxicology* 17: 362–371
76. Yoon K-Y, Hoon Byeon J, Park J-H, Hwang J (2007) Susceptibility constants of *escherichia coli* and *bacillus subtilis* to silver and copper nanoparticles. *Sci Total Environ* 373:572–575
77. Mallick S, Sharma S, Banerjee M, Ghosh SS, Chattopadhyay A, Paul A (2012) Iodine-stabilized Cu nanoparticle chitosan composite for antibacterial applications. *ACS Appl Mater Interfaces* 4:1313–1323
78. Karlsson HL, Cronholm P, Gustafsson J, Möller L (2008) Copper oxide nanoparticles are highly toxic: a comparison between metal oxide nanoparticles and carbon nanotubes. *Chem Res Toxicol* 21:1726–1732
79. Hu W, Peng C, Luo W, Lv M, Li X, Li D, Huang Q, Fan C (2010) Graphene-based antibacterial paper. *ACS Nano* 4:4317–4323
80. Paret LM, Vallad EG, Averett RD, Jones BJ, Olson MS (2013) Photocatalysis: effect of light-activated nanoscale formulations of TiO₂ on *xanthomonas perforans*, and control of bacterial spot of tomato. *Phytopathology* 103:228–236
81. Panáček A, Kvítek L, Prucek R, Kolář M, Večeřová R, Pizurová N, Sharma VK, Nevěčná T j, Zbořil R (2006) Silver colloid nanoparticles: synthesis, characterization, and their antibacterial activity. *J Phys Chem B* 110:16248–16253

82. Xiu Z, Zhang Q, Puppala HL, Colvin VL, Alvarez PJJ (2012) Negligible particle-specific antibacterial activity of silver nanoparticles. *Nano Lett* 12:4271–4275
83. Xu W-P, Zhang L-C, Li J-P, Lu Y, Li H-H, Ma Y-N, Wang W-D, Yu S-H (2011) Facile synthesis of silver@graphene oxide nanocomposites and their enhanced antibacterial properties. *J Mater Chem* 21:4593–4597
84. Das MR, Sarma RK, Saikia R, Kale VS, Shelke MV, Sengupta P (2011) Synthesis of silver nanoparticles in an aqueous suspension of graphene oxide sheets and its antimicrobial activity. *Colloids Surf B Biointerfaces* 83:16–22

Current Status and Future Prospects of Next-Generation Data Management and Analytical Decision Support Tools for Enhancing Genetic Gains in Crops



Abhishek Rathore, Vikas K. Singh, Sarita K. Pandey, Chukka Srinivasa Rao, Vivek Thakur, Manish K. Pandey, V. Anil Kumar, and Roma Rani Das

Abstract Agricultural disciplines are becoming data intensive and the agricultural research data generation technologies are becoming sophisticated and high throughput. On the one hand, high-throughput genotyping is generating petabytes of data; on the other hand, high-throughput phenotyping platforms are also generating data of similar magnitude. Under modern integrated crop breeding, scientists are working together by integrating genomic and phenomic data sets of huge data volumes on a routine basis. To manage such huge research data sets and use them appropriately in decision making, Data Management Analysis & Decision Support Tools (DMASTs) are a prerequisite. DMASTs are required for a range of operations including generating the correct breeding experiments, maintaining pedigrees, managing phenotypic data, storing and retrieving high-throughput genotypic data, performing analytics, including trial analysis, spatial adjustments, identifications of MTAs, predicting Genomic Breeding Values (GEBVs), and various selection indices. DMASTs are also a prerequisite for understanding trait dynamics, gene action, interactions, biology, GxE, and various other factors contributing to crop improvement programs by integrating data generated from various science streams. These tools have simplified scientists' lives and empowered them in terms of data storage, data retrieval, data analytics, data visualization, and sharing with other researchers and collaborators. This chapter focuses on availability, uses, and gaps in present-day DMASTs.

A. Rathore (✉), V. K. Singh, S. K. Pandey, C. S. Rao, V. Thakur, M. K. Pandey, V. Anil Kumar, and R. R. Das
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India
e-mail: a.rathore@cgiar.org

Graphical Abstract



Keywords Analytical Decision Support Tool, Data management, Genetic gains, Plant breeding

Contents

1	Introduction	278
2	DMAST for Phenotypic Evaluation of Datasets	280
3	DMAST for Molecular Marker Datasets Including Genomics Data	281
4	DMAST for Metabolomics and Proteomics Data	283
5	DMAST for Molecular Breeding	284
6	Integrated Pipelines for Plant Breeding Data Management	285
7	DMASTs for Data Sharing and Visualization	287
8	Breeder Requirements for Enhancing Genetic Gains	288
	8.1 Pipeline to Understand the Association Between Phenotype and Genotype	288
	8.2 High-Throughput and Precision Phenotyping	289
	8.3 New Web-Based Interface with Better Organization	289
	8.4 Trait Ontology Inference as Part of the Data Management Pipeline	289
	8.5 Better Support from Plant Genomics	290
	8.6 Better Support for Data Analysis and Investments	290
	8.7 Integration from “Omics” Information	290
9	Conclusion	290
	References	291

1 Introduction

Good quality research experiments, precise data, appropriate data analysis, and data-driven decision making make up the backbone of modern agricultural research and integrated breeding. Integrated breeding exploits high-throughput phenomics and genomics, and has opened the floodgates of data pouring into crop specialists of all disciplines. Many international consortia and research centers are engaged in plant

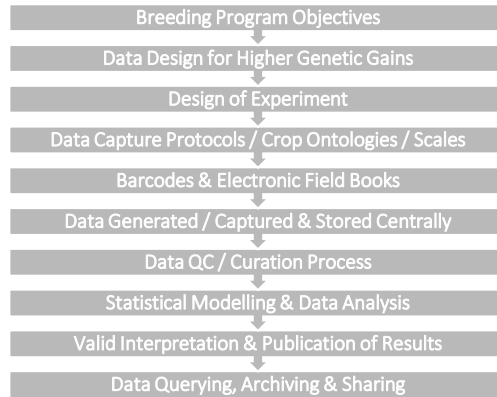


Fig. 1 Data management pipeling of a successful breeding program. Traditionally, data management has been interpreted as storing research data in online repositories and share. However, data management has a much wider definition, which starts with designing the study and concludes with appropriate analysis and making the data publicly available through repositories

research and are generating huge amounts of sequencing/genotyping and phenotyping data. These data sets require the appropriate capacities for processing, analysis, management, and storage. Often it becomes very difficult to analyze these data sets and convert them into information through conventional data-management tools and analysis strategies. It has also been observed that researchers very frequently limit the definition of data management and interpret it in terms of mere physical data storage and access. However, the scientific data management scope is much wider and includes a complete life cycle. Figure 1, explains a typical data-management workflow in a breeding pipeline.

Usually databases and analytical tools are required for efficient utilization, retrieval, analysis, and decision making at each step of the genetic gain-enhancement process. Considering this need, experts in the area of bioinformatics, biometrics, and statistical genomics, in collaboration with plant geneticists, have developed many software tools and protocols for analyzing the data. Nevertheless, there is still tremendous scope for developing more efficient and user-friendly analysis and decision making to speed up the process of achieving higher genetic gain in crop plants.

The need for informatic intervention is required at all the steps of integrated breeding, such as selection of the appropriate experimental design, determining the size of the population, modern ways of data collection, use of modern databases, BLUE/BLUP with spatial adjustments being made during phenotypic selections, enabling sample tracking for DNA sample collections, genetic map construction, population structure, identification of marker-trait association, background and foreground selection, combining favorable alleles for complex traits using marker-assisted recurrent selection (MARS), and estimating genomics estimated breeding values (GEBVs) in genomic selection (GS). There are various open-source and proprietary tools available that cater to each step discussed above. Selection of these tools varies according to the hardware requirements, operating system, the degree of computer

skills required, user-friendliness, statistical models and algorithms used for analysis, corroboration of input data, and visualization of output results.

Development of integrated pipelines combining various useful software in one place also played a major role in the efficient integrated breeding program. Several such analytical pipelines are available that combine the analysis of phenotypic and genotypic data, such as running mixed models for statistical analysis, construction of linkage maps, mapping of quantitative traits, GWAS, and GS, etc. Nevertheless, some important DMASTs have been developed not only for helping with integrated breeding, but are also helpful in managing a larger set of data management for a future breeding program. We discuss here several informatics tools, data sharing and visualization platforms, their comparative usefulness experienced by researchers, and their ease of use, popularity, and prospects of improvement with regard to current technological needs and statistical methods.

We present several DMASTs in different sections, which include (a) DMASTs for phenotypic evaluation of datasets, (b) DMASTs for molecular marker datasets, (c) DMASTs for metabolomics and proteomics data, (d) DMASTs for molecular breeding, (e) integrated pipelines for plant breeding data management, and (f) DMASTs for data sharing and visualization. The last section is on breeder requirements for enhancing genetic gain.

2 DMAST for Phenotypic Evaluation of Datasets

High-throughput phenotyping generates a large volume of different types of data including nominal, categorical, ordinal, and ratio types of data sets. To capture variation and make good interpretations out of generated datasets, data should be subjected to appropriate statistical techniques. Good analysis will only be possible if the study was designed keeping the hypothesis in mind and data were also subjected to appropriate quality checks. The data must be cleaned, curated, and well summarized before final analysis and interpretation of results.

There is a range of statistical analysis available for the analysis of agricultural experiment data. A description of all of these is not possible in a single chapter and also out of the scope of this book, but we recommend appropriate selection of random and fixed factors and the use of mixed models, possibly with spatial adjustments. It is worth mentioning that analysis should be performed by standard and well-known statistical packages. Several commercial, open-source, and free software systems for statistical analysis are available. Often the commercial software is expensive, whereas the majority of the freeware has limited functions or is sometimes difficult to use. Among the commercially available software, ASREML (<https://www.vsnl.co.uk>), Genstat (<https://www.vsnl.co.uk/software/genstat/>), MINITAB, Statistical Package for Social Sciences (SPSS; <http://www.spss.co.in/>), Statistical Analysis System (SAS; <https://www.sas.com>), Statistica (<https://software.dell.com/products/statistica/>), and STATA (www.stata.com/) are very common, relatively easy to use, and can perform most data analyses and visualization for making breeding decisions.

There are plenty of free and open-source tools that are also available for performing statistical data analysis. R (<https://www.r-project.org/>) and Python (<https://www.python.org/>) are two such environments for performing sound statistical computing and visualization. These languages have become increasingly popular due to their versatility and availability of sound programming environments similar to many commercially available environments. R has a wide community support and has a core component that implements many classical and modern statistical methods. One can build on top of core functionality, develop their own code, and pack in R packages to perform customize analysis. The benefit of community support is that many analyses that are not available as part of core functions are also available to end users. R can also be interfaced with other programming languages and GUI development tools, such as Galaxy, Java, and Tk/Tcl. PBTools and CropStat (<http://bbi.irri.org/products>) are such free applications for plant breeders. Under open-source statistical software, R has emerged as the leader and most important software for analyzing data from all agricultural disciplines.

Python on other hand is also gaining popularity, but it seems it will take some more time to gain a strong place in the academic world. The advantages of Python are an easy learning curve and good graphics and visualization capabilities. Python has been used widely in web development and hence the development of online web-based analytical applications is a clear advantage with Python.

3 DMAST for Molecular Marker Datasets Including Genomics Data

To analyze genetic diversity with a moderate level of markers and genotypes, NTSYSpc (numerical taxonomy and multivariate analysis system) [1] is one of the most widely used programs, as is evident from the citation index. MEGA7 (Molecular Evolutionary Genetic Analysis) is another highly cited and widely used program, and was originally developed in 1993 [2]. This software can estimate the evolutionary distance or the phylogenetic tree calculation of genetic distance, using DNA or protein sequences data. This is a flexible and easy-to-use genetic data analysis system, and it can import unlimited sizes of datasets from various programs. DARwin (<http://darwin.cirad.fr/>) is a freely available software package developed for diversity and phylogenetic analysis by evolutionary dissimilarities. DAMBE (data analysis for molecular biology and evolution) is a phylogenetic analysis software first released in 2001 and recently updated as DAMBE5, with many new functions [3]. PAUP (phylogenetic analysis using parsimony) (<http://paup.csit.fsu.edu/>) is another widely-used program for inferring and interpreting evolutionary trees. It includes analysis of parsimony, distance matrix, invariance and maximum likelihood methods, and other statistical analysis.

To analyze population genetics, GENEPOP is a widely used population genetics software [4]. This software can estimate the number of tests (null allele estimates,

exact tests, Markov chain probabilities, and test statistics), multi-locus F-statistics, microsatellite allele sizes, RST, and rST, etc. Arlequin [5] is highly cited software for analyzing population genetics, and this software can handle a large number of datasets including molecular variance in the population regarding AMOVA, which is the unique feature of this software. Power Marker [6] is a program designed for SSR or SNP marker data for population genetic analysis, with a user-friendly graphic interface. DnaSP v5 (DNA Sequence Polymorphism) [7] is another software package for the analysis of nucleotide polymorphism from aligned DNA sequence data. SMOGD (Software for the Measurement of Genetic Diversity) [8] is a web-based application for the calculation of advanced proposed genetic diversity indices $G'ST$ and $Dest$. GenAIEx 6.5 [9] is a Microsoft Excel-based software, and it offers a wide range of population genetic analysis options for the full spectrum of genetic markers.

Genome-mapping methods such as the construction of a genetic/linkage map and a physical map make up one of the basic steps involved in the identification of genes/QTLs for the trait of interest in the target environment. Mapping involves some steps such as determining recombination fractions, using a mapping function (Haldane or Kosambi), testing for appropriate linkages (LOD scores), grouping and ordering of markers into linkage groups, and bridging different genetic maps to develop a consensus map. Within the toolkit available for this work, MAPMAKER [10] open-source software released during the 1980s led the way towards computational strategies in the construction of the genetic map. This software uses a multipoint likelihood objective function [11] by combining the EM algorithm and the Hidden Markov Model (HMM) method, which significantly lowers the computational time when large datasets are used for analysis. The MAPMAKER software is still quite popular among geneticists, as the paper from Lander et al. [10] shows, with more than 6,000 citations in <http://scholar.google.co.in/>. However, due to the command prompt interface, it is not very user-friendly, and good quality graphic representation cannot be generated using it. JoinMap [12] is a widely used software as it has several positive features such as the user-friendly MS-Windows interface, ability to integrate maps from different mapping populations, continuous development, and professional support. JoinMap utilizes maximum likelihood and regression mapping algorithms for marker order strategies. For a better graphic representation, MapChart [13] and cMap [14] are also quite popular tools. For handling large numbers of marker datasets, special software packages have recently been developed, such as MadMapper [15] and MSTmap [16], for making a high-density genetic map.

Once the genetic map is made, the next step is to identify marker-trait associations by QTL analysis. As most agronomically important traits/phenotypes are polygenic in nature, many statistical and genetic models have been developed. MapMaker/QTL [17] was one of the most widely used open-source software during the 1990s and utilizes interval mapping (IM). As it is a command prompt-based interface and does not handle complex statistical models such as multiple interval mapping (MIM) and composite interval mapping (CIM), it is currently not much in use. QTL Cartographer [18] and QGene [19] do both MIM and CIM analysis. Software IciMapping [20] has a better QTL analysis model called inclusive composite interval mapping (ICIM). Recently, a new software called QTLnetwork [21] is becoming quite popular among geneticists as

it can analyze all types of genetic models such as additive QTLs, additive and epistatic QTLs, and QTL \times environment interactions [22].

To understand the germplasm, STRUCTURE (<http://pritchardlab.stanford.edu/structure.html>) is the most extensively used software to detect population genetic structure. This program generates clusters caused by admixture between populations [23]. EIGENSOFT and Bayesian Analysis of Population Structure (BAPS) are the two another widely used statistical packages for detection and correction of population stratification in GWAS analysis [24, 25]. A detailed list of other available software packages for linkage disequilibrium analysis can be found in the following link: <http://www.genes.org.uk/software/LD-software.shtml>. Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL) is the most common and highly cited software for performing marker-trait association analysis in GWAS studies in plants [26]. PLINK is another highly cited open-source whole genome association analysis toolset, which performs a range of basic, large-scale analyses in a computationally efficient manner [27].

4 DMAST for Metabolomics and Proteomics Data

In addition to genomics, proteomics and metabolomics hold a great perspective for serving as pillars for crop improvement. Complex and multi-omics studies have increased in the recent past, which integrate genomics data with metabolomics, epigenetics, and proteomics data. It is anticipated that in the future metabolomics will emerge as a significant part of crop improvement programs for achieving complex breeding objectives. Therefore, metabolomics techniques will be integrated with other “omics” technologies in order to identify and understand biochemical mechanisms and their consequences [28]. The few commonly used software programs available are BioCyc (<http://biocyc.org>), iPath (<http://pathways.embl.de>), KaPPA-View (<http://kpv.kazusa.or.jp/en/>), KEGG (<http://www.genome.jp/kegg/pathway.html>), MapMan (<http://mapman.gabipd.org/web/guest/mapman>), MetabolomeExpress (<https://www.metabolome-express.org/>), MetaboAnalyst (<http://www.metaboanalyst.ca/faces/home.xhtml>), Metscape (<http://metscape.ncibi.org>), MGV (<http://www.microarray-analysis.org/mayday>), Paintomics (<http://www.paintomics.org>), Pathos (<http://motif.gla.ac.uk/Pathos/>), Pathvisio (<http://www.pathvisio.org/>), PRIME (<http://prime.psc.riken.jp/>), ProMetra (http://www.cebitec.uni-bielefeld.de/groups/brf/software/prometra_info/), Reactome (<http://www.reactome.org>), VANTED (<http://vanted.ipk-gatersleben.de>), and MetPA (<http://metpa.metabolomics.ca>). Most computational tools available are largely intended for metabolite identification. However, in order to gain some biological insight, it is necessary to have an integrated tool that can perform metabolite identification, functional analysis, detection of associated compounds, and metabolic modeling [28]. At the same time, there has been a swift addition of proteomics data due to advances in proteomics technologies such as high-throughput experimental

platforms [29]. There are a number of data repositories as well as data analysis and visualization tools available for proteomics [30–32].

PRoteomicsIDentifications database (PRIDE; <http://www.ebi.ac.uk/pride>) is a comprehensive database of protein and peptide identifications; MSDA (<https://msda.unistra.fr/>) is a proteomics suite for detailed Mass Spectrometry Data Analysis; COMPASS (<https://github.com/dbaileychess/Compass>) is a suite of pre- and post-search proteomics software tools for OMSSA; PICR (<http://www.ebi.ac.uk/Tools/picr/>) or CRONOS [33] are web-based algorithms that associate names of the protein with their corresponding gene names; Gene Ontology terms (<http://www.geneontology.org>) are used to connect the protein identifier with its associated Gene. Obtained MS/MS spectra are interpreted with Mascot (<http://www.matrixscience.com>) and SEQUEST (<https://omictools.com/sequest-tool>) algorithms. Some functional databases such as the “Uniprot knowledge base” (www.uniprot.org/help/uniprotkb) and Ensembl (www.ensembl.org/) are being widely used in the field of proteomics along with other detailed pathway databases like KEGG (www.genome.jp/kegg/pathway.html), Reactome (<http://www.reactome.org>), and Ingenuity Pathway Knowledge Base (<https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/>). In addition to comprehensive resources, precise databases have been established for signal transduction processes, such as PANTHER (<http://pantherdb.org/about.jsp>). Information on protein interactions in complexes are deposited in interaction databases such as BioGRID (<https://thebiogrid.org/>) and IntAct (<http://www.ebi.ac.uk/intact>). Further, STRING (<https://string-db.org/>) and Cytoscape (www.cytoscape.org/) are graphic tools for visualizing and analyzing biological pathways. EnrichNet (www.enrichnet.org/) serves as a web-based platform, integrating pathway and interaction analysis in several databases (KEGG, Gene Ontology, Reactome, Wiki, and NCI pathways (<http://www.wikipathways.org/index.php/WikiPathways>)). A few other programs like Pfam (<http://pfam.xfam.org/>), Interpro (<https://www.ebi.ac.uk/interpro/>), SMART (<http://smart.embl-heidelberg.de/>), and DAVID (<https://david.ncifcrf.gov/>) are among the commonly used software programs.

Therefore, in the future there is a need for an integrated tool to support the analysis and interpretation of multi-omics data generated from different fields consisting of large populations. Development and deployment of the DMASTs for metabolomics and proteomics in accordance with the necessity of breeding programs will help to achieve breeding targets efficiently and rapidly.

5 DMAST for Molecular Breeding

Once the genomic region has been identified through QTL analysis, these regions are then introgressed/pyramided into elite cultivars through the marker-assisted backcrossing (MABC) approach. To quickly introgress the targeted genomic regions, strategies such as foreground selection, recombination selection, and recovery of recurrent parent genome (RPG) through background selection are utilized.

Several visualization tools have been developed in the past such as GGT (graphical genotype) [34] and Flapjack [35], and are currently being used alone or as part of pipelines such as iMAS (<http://www.icrisat.org/bt-biometrics-imas.htm>) and ISMU (Integrated SNP Mining and Utilization) [36]. The Marker-assisted Back-crossing Tool (MABT) is another JAVA-based decision-making software program that enables users to calculate the percentage of recovery of the recurrent parent at each generation (<https://www.integratedbreeding.net/ib-tools/breeding-decision/marker-assisted-back-crossing-tool>). To implement MARS in the breeding program through accelerated genetic gain by assembling favorable alleles issued from diverse parents, OptiMAS [37] has been developed with the following interactive graphical interface: (a) to trace parental alleles throughout generations, (b) to select the best plants based on estimated molecular scores, and (c) for an efficient inter-mating strategy to recombine positive alleles in a single genetic background. Genomic selection (GS) is a new molecular breeding approach using whole-genome profiling with a large number of markers and offers many advantages involved with improving the rate of genetic gain in crop breeding programs. solGS [38] and ISMU2 are two programs available for the calculation of GEBVs for the selection of individuals.

6 Integrated Pipelines for Plant Breeding Data Management

Data management plays a major role in creating a basis for sound scientific decision making, increased efficiency of resource use, and ultimately leads to enhanced research quality and reliability [39]. Data management software is not just a database but signifies appropriate experimental design, analysis, interpretation, archiving, and sharing of data. One of the biggest challenges for effective data management in public plant breeding is a lack of access to public data management systems to track samples, manage and analyze breeding data, and support breeding decisions. To overcome this hindrance, a few commercial software programs have been developed that offer breeding management systems; however, these come with an additional cost to the research organizations. Intensive crop improvement data demands a single integrated platform that can be used for data management, data mining, analysis, and sharing.

Many attempts have been made from both public and private sectors to provide advanced systems for data management. However, some of the systems have multiple features while others have specific applications [40]. Most importantly, the DMASTs need to evolve with the pace of volume and type of data generated in fast-evolving genetic and breeding methodologies. For this reason, currently no single data-management tool can be used for all the applications. Nevertheless, the scientific community is now well aware of such a need and soon there will be a few initiatives to work in this direction, for example, the development of the International Crop Information System (ICIS) (www.icis.cgiar.org) by the CGIAR and partners, a database system for the management and integration of global information on crop improvement and genetic resources for any crop [41].

To efficiently manage the regular movement of data from lab to the breeder and to integrate information from genotyping and phenotyping, comprehensive crop-improvement data-management tools are required. To deal with the constraints in present-day data management, the Integrated Breeding Platform (IBP) (<http://www.integratedbreeding.net>), established by the CGIAR's Generation Challenge Program (GCP) and partners, offers a web-based frontline platform of technology and services for managing both traditional and modern breeding activities. From phenotyping to complex genotyping, it provides information, analytical tools, and related services to conduct modern breeding research. The Breeding Management System (BMS) of the IBP is an interconnected application specifically designed for managing breeding activities through all phases of research using various types of data management, statistical analysis, and decision support tools. Presently, the BMS is the only publicly available data management solution that supports various crops and has inbuilt international crop ontologies. The BMS is actively used by many CGIAR institutes including ICRISAT, CIAT, and IITA, with many more institutes adopting it. ICRISAT is one of the first centers to adopt it on an institutional scale and to implement it on a cloud. The BMS has an advantage of hosting several crops on one installation.

Breeding4Rice (B4R), is a breeding information management system at IRRI that provides an integrated, user-friendly information management system, developed using modern web technologies, and is deployed to a cloud infrastructure. The system is being extended to various other crops and will soon be available for maize and wheat. CassavaBase (<https://www.cassavabase.org/>) is an integrated information management system for breeding programs that deals with phenotyping, low-density marker, pedigree management, and selection decision support. Katmandoo (<http://www.katmandoo.org/>) is a data management system of biosciences primarily developed to be used by breeders and researchers in breeding programs. It is mainly focused on providing single tools for dealing with both phenotypic and genotypic data.

In addition to the above free and open-source databases, there are several commercial software solutions that are also available for handling the breeding data pipeline. As all systems are at the same stage of development, no clear-cut comparisons of these software programs are available. However, the authors of this chapter have experience in using a couple of them, and one major drawback that we observed is that once the user stops paying the annual renewal fee, there is no way one can even log in to the system and work with their past experiments. The first tool in this line is PRISM, a plant-breeding software solution (<http://www.teamcssi.com/index.html>) for plant researchers and agronomists. It provides user-friendly tools to manage breeding data. PRISM has been used by various public and private breeding institutions and is known for its flexible architecture. Another popular data pipeline is the Phenome One platform (<http://phenome-networks.com/solutions/for-plant-breeders/>). This platform supports all stages of the breeding process for field crops, horticulture crops, and ornamental plants. It is a web-based and user-friendly system, and also supports data analytics and integrated mobile application. Similarly, AGROBASE Generation II (<http://www.agronomix.com>) is a Windows-based agronomy software system. The CORE System

of AGROBASE Generation II offers data management and analytical tools for crop improvement. Progeny (<http://www.progeno.net/software>) is a Ghent University spin-off company that aims to empower plant breeders by providing access to breeding and selection methods. Several other platforms include Progen software (<http://www.progeno.net/>), which permits plant breeders to improve selection efficiency by incorporating phenotyping and genotyping data in the decision process. E-Brida (<http://www.agripartner.nl/en-us/products/plantbreedingsoftware.aspx>) is a breeding information system with several options for data recording and analysis. GeneFlow (<http://www.geneflowinc.com>) is a software program that provides a comprehensive tool for integrating pedigree, phenotype, and genotype data.

7 DMASTs for Data Sharing and Visualization

Research data are extremely valuable assets and resources, and good management of research data is essential for research excellence. It is essential to facilitate data sharing and ensure the sustainability and accessibility of data in the long term, and thus, their re-use for future science. This permits new and innovative research to be built on existing information, which is especially true for cases where public investment in research is to be realized. With well-organized and accurate research data we can get high quality research outputs and scientific discoveries based on evidence, while using less resources. With good data management practices and proper planning, researchers can benefit greatly, especially in saving cost and time.

Currently, many funding agencies ask for consideration of open data and data sharing for all research projects they fund, and impose research data requirements that focus on how data will be preserved and shared for public use after the project is completed. Scientific data have very important value beyond their use for the original research. Data sharing and visualization encourages scientific enquiry and debate, and promotes innovation, which may lead to new collaborations between data users and data authors, enhances the impact and visibility of research, can provide a direct credit to the researcher as a research output, and promotes the research that created the data and its outcomes. A critical part of making data findable, accessible, interoperable, and reusable with long-lasting usability is to ensure that it can be interpreted and understood by any user even in the future.

Several open-source tools are available for effective and efficient data sharing with different capacities. Data sharing helps in the reuse of existing data for new studies, which can result in innovations and new opportunities. There are many open-source data management tools available that can be used at an institute or project level. Dataverse (<https://dataverse.harvard.edu/>) is a research data storage and sharing platform developed by Harvard University, Cambridge, MA, USA, which is freely downloadable and can establish its own institutional open data repository. This platform is well integrated with R software modules and Geospatial map generation. Several CGIAR institutions have implementations of Dataverse and are using it as their primary data-sharing software. Dataverse is highly configurable

and can be queried through well-defined APIs. CKAN (<https://ckan.org>) is also an open-source data portal and data management solution that provides a streamlined way to make data discoverable and presentable with a rich collection of metadata, making it a valuable and easily searchable data catalog. Researchspace (<https://www.researchspace.com>) is a research management tool for Principal Investigators (PIs) and research team members of specific groups, to observe and manage lab workflows, capture, archive, organize, publish, and share the data. e!DAL (<https://edal.ipk-gatersleben.de/>) is a lightweight software framework for publishing and sharing research data, the main features being: version tracking, metadata management, information retrieval, an embedded HTTP(S) server for public data access, access to a network file system, and a scalable storage backend. DSpace (<http://www.dspace.org>) is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. DSpace preserves an open access format for all types of digital content. Usually open access repositories are used for publishing digital content with more focus on long-term storage, access, and preservation. Fedora (<http://fedorarepository.org>) is a robust, modular, open-source repository system for the management and dissemination of digital content. It is especially suited for digital libraries and archives, for both access and preservation.

8 Breeder Requirements for Enhancing Genetic Gains

Enhancing genetic gains for crop improvement demanded several automated, integrated, straightforward, and easy to use pipelines. Based on several reports and publications, we have listed a few of the essential requirements from the breeders' perspective, which includes: (a) a pipeline to understand associations between phenotype and genotype, (b) high-throughput precision phenotyping, (c) a new web-based interface with better organization, (d) a trait ontology function inference as part of the data management pipeline, (e) better support from plant genomics, (f) better support for data analysis, and (g) integration from "omics" information. Based on the above requirements, the breeding pipeline should have a seamless interconnected analytical solution for different applications in crop improvement.

8.1 *Pipeline to Understand the Association Between Phenotype and Genotype*

The central challenge of modern data management tools are weak genomics to phenomics links. This also highlights the need for careful pipeline development and advocates for the inclusion of a robust and straightforward platform that can correlate between phenomics and genomics data seamlessly. For example, several CG centers (3,000 rice accessions from IRRI, Philippines, and 3,000 chickpea accessions from ICRISAT, Hyderabad) have generated a huge amount of genotyping/re-sequencing data. Multi-location phenotyping

data of such lines will provide meaningful results to the breeders if simple-to-use pipelines are available for understanding the association between phenotype and genotype. There exist pipelines that do part of this job and do not cycle through start to end. The current need is to bring efficiency to these tools and to link them to each other in order to undertake the huge phenotypic and genotypic datasets generated in breeding programs.

8.2 High-Throughput and Precision Phenotyping

The emphasis on high-throughput and precision phenotyping represents a significant change for breeders engaged in variety development who have traditionally favored simplicity, speed, and flexibility over sensitivity, precision, and accuracy. This is because historically the advantages of the latter could not be translated into an economically relevant genetic gain in a breeding context, and this is why easy, fast, and efficient phenotyping-capturing tools are the need at present. For example, PHENOME, Field Book, 1KK, and Coordinate are recent high-throughput phenotyping, software programs/Android apps that allow researchers to accumulate, categorize, and manage a large volume of phenotypic data using Android smartphones with barcode scanners or a Personal Digital Assistant (PDA) with a built-in barcode scanner. The collected data in the smart device could be easily transferred for data analysis in any operating system through the appropriate DMAST.

8.3 New Web-Based Interface with Better Organization

Many of the DMASTs or data management tools are stand-alone and can only be utilized through better infrastructure and with high IT skill manpower. Therefore, in the near future cloud-based, simple-to-use tools are required for breeders, which could be utilized on simple PCs. An advantage of such a web-based system will be that such tools can be used from any place or PC through a simple login with a user ID and password. The other major advantage of such a system is that the huge submitted phenotypic/genotypic datasets will be safer than those saved on standalone PCs.

8.4 Trait Ontology Inference as Part of the Data Management Pipeline

Trait ontology function should be an integral part of the data-management pipeline. This will be useful for the selection of diverse lines, for making new crosses, or for the development of new combinations of hybrids. This feature will be helpful for understanding the contributions of diverse parents in breeding lines, through their performance.

8.5 *Better Support from Plant Genomics*

Another important requirement from the breeders' perspective is better support from plant genomics scientists in the identification of trait-associated markers for complex traits, the selection of which is difficult in field conditions. Additionally, the development of a purity kit is important, not only for the purity of parental lines and hybrids but also for high-yielding varieties, so that the seed purity of the lines/varieties/hybrids can be tested in less time. Better GS prediction models with high prediction accuracy will also be useful for breeders for enhancing genetic gains through genomics interventions.

8.6 *Better Support for Data Analysis and Investments*

Meaningful and timely data analysis is the critical component of breeders' success in enhancing genetic gain. Most of the breeding trials and genotype-to-phenotype correlation requires specific DMASTs, and it is sometimes difficult for the breeders to use these tools in their breeding programs with limited infrastructure. Therefore, breeders require professional data analysis for analyzing complex datasets with specifically required tools. The information provided by such analysis of these huge datasets will be useful for making critical decisions in breeding programs. There is a need for strengthening investment in data analysis in breeding programs.

8.7 *Integration from “Omics” Information*

Besides genomics and phenomics, multiple studies have been conducted in other “Omics” fields in many crop plants. These “Omics” studies include transcriptomics, epigenomics, proteomics, and metabolomics. They will develop a better understanding of traits and generate meaningful information that can be used during plant selection in the field. Such integration of this information with DMASTs will increase the precision of decision making in plant selection.

9 Conclusion

This chapter discusses the status and future prospects of next-generation data management and analytical and decision support tools for crop improvement. We have presented a critical appraisal of different DMASTs and data management tools along with integrated pipelines. We have also presented the breeders' future requirements for enhancing genetic gains in terms of new required tools and easy-to-use

pipelines. We believe that the availability of GUI-based platforms with appropriated DMASTs will help breeders to make the best use of these tools in their breeding programs. Development and deployment of the right DMASTs at the right time will usher the crop improvement programs into a modernized knowledge-based crop improvement era towards sustainable crop production.

References

1. Rohlf FJ (1992) NTSYS-pc: numerical taxonomy and multivariate analysis system. Appl Biostat, ISBN 9780925031181
2. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30:2725–2729
3. Xia X (2013) DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol* 30:1720–1728
4. Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* 86:248–249
5. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinforma* 1:47–50
6. Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21(9):2128–2129
7. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11):1451–1452
8. Crawford NG (2010) SMOGD: software for the measurement of genetic diversity. *Mol Ecol Resour* 10(3):556–557
9. Peakall PE, Smouse R (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28:2537–2539
10. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174–181
11. Lander E, Green P (1987) Construction of multilocus genetic maps in humans. *Proc Natl Acad Sci U S A* 84:2363–2367
12. Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J* 3:739–744
13. Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. *J Hered* 93:77–78
14. Fang Z, Polacco M, Chen S, Schroeder S, Hancock D, Sanchez H, Coe E (2003) cMap: the comparative genetic map viewer. *Bioinformatics* 19:416–417
15. Kozik A, Michelmore R (2006) MadMapper and CheckMatrix-python scripts to infer orders of genetic markers and for visualization and validation of genetic maps and haplotypes. In: Proceedings of the plant and animal genome XIV conference, San Diego. Abstract P957/CP013-http://www.intl-pag.org/14/abstracts/PAG14_C013.html
16. Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet* 4:e1000212
17. Lander ES, Bostein DR (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage map. *Genetics* 121:185–189
18. Basten CJ et al (1994) Zmap—a QTL cartographer. In: Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W, Gibson JP, Kennedy BW, Burnside EB (eds) Proceedings of the 5th World Congress on genetics applied to livestock production: computing strategies and software, vol 22. The organizing committee, 5th World Congress on genetics applied to livestock production, Guelph, pp 65–66

19. Nelson JC (1997) QGENE: software for marker-based genomic analysis and breeding. *Mol Breed* 3:239–245
20. Li H, Ye G, Wang J (2007) A modified algorithm for the improvement of composite interval mapping. *Genetics* 175:361–374
21. Yang J, Hu C, Hu H, Yu R, Xia Z, Ye X, Zhu J (2008) QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. *Bioinformatics* 24:721–723
22. Su C, Qiu X, Ji Z (2013) Study of strategies for selecting quantitative trait locus mapping procedures by computer simulation. *Mol Breed* 31:947–956
23. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multi-locus genotype data. *Genetics* 155:945–959
24. Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinform* 9(1):539
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
26. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575
28. Chagoyen M, Pazos F (2013) Tools for the functional interpretation of metabolomic experiments. *Brief Bioinform* 14:737–744
29. MacBeath G (2002) Protein microarrays and proteomics. *Nat Genet* 32:526–532
30. Falkner JA, Ulintz PJ, Andrews PC (2006) A code and data archival and dissemination tool for the proteomics community. *Am Biotechnol Lab* 24(5):28
31. Vizcaíno AJ, Côté RG, Csordas A, Dianas JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J et al (2013) The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 41:D1063–D1069
32. Wein SP, Côté RG, Dumousseau M, Reisinger F, Hermjakob H, Vizcaíno AJ (2012) Improvements in the protein identifier cross-reference service. *Nucleic Acids Res* 40:276–280
33. Waegelé B, Dunger-Kaltenbach I, Fobo G, Montrone C, Mewes HW, Ruepp A (2009) CRONOS: the cross-reference navigation server. *Bioinformatics* 25:141–143
34. vanBerloo R (2008) GGT 2.0: versatile software for visualization and analysis of genetic data. *J Hered* 99:232–236
35. Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WT, Flavell AJ, Marshall D (2010) Flapjack-graphical genotype visualization. *Bioinformatics* 26:3133–3134
36. Azam S, Rathore A, Shah TM, Telluri M, Amindala B, Ruperao P et al (2014) An integrated SNP mining and utilization (ISMU) pipeline for next generation sequencing data. *PLoS One* 9: e101754
37. Valente F, Gauthier F, Bardol N, Blanc G, Joets J, Charcosset A, Moreau L (2013) OptiMAS: a decision support tool for marker-assisted assembly of diverse alleles. *J Hered* 104:586–590
38. Teclé IY, Edwards JD, Menda N, Egesi C, Rabbi IY, Kulakow P, Mueller LA (2014) solGS: a web-based tool for genomic selection. *BMC Bioinform* 15(1):398
39. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M (2011) Data sharing by scientists: practices and perceptions. *PLoS One* 6:e21101
40. Xu Y (2010) *Molecular plant breeding*. CAB International, Nosworthy Way
41. McLaren CG, Bruskiewich RM, Portugal AM, Cosico AB (2005) The international rice information system. A platform for meta-analysis of rice crop data. *Plant Physiol* 139:637–642

Index

A

Abiotic stress, 7, 65, 118, 161, 169–178, 187, 213, 219, 241, 243
 miRNAs, 176
 siRNAs, 178
Accumulation, 69, 148, 178, 193, 200, 210–222
Acibenzolar-*S*-methyl (ASM), 269
Adapters, 28, 40
Advanced backcross QTL (AB-QTL), 68
Advanced intercross (AIC), 121, 133
Adzuki bean (*Vigna angularis*), 28
Alleles, 20, 30, 41, 65, 67, 69, 117–131, 142, 150, 256, 279, 281, 285
Amino acids, 210–221
Analytical decision support tools, 8, 187, 277
Antimicrobial properties, 263, 270
Antioxidants, 196, 211, 215, 218, 220, 264, 270
Arabidopsis Genome Initiative (AGI), 17
Arabidopsis multi-parent recombinant inbred line (AMPRIL), 110, 131
Ascorbate glutathione (GSH) cycle, 215
AS regulatory proteins, 166
Assembly, 11, 15
 strategies/technologies, 29
Association mapping (AM), 6, 56, 62, 83, 86, 115, 130

B

BACs, libraries, 17
Bacteria, 8, 14, 175, 192, 220, 263
Barley (*Hordeum vulgare*), 40, 42, 61, 67–69, 87, 101, 124, 127, 147, 190, 214, 218, 219

Bayesian analysis/methods, 62, 88, 131
Bayesian Analysis of Population Structure (BAPS), 283
Bayesian interval mapping (BIM), 57
Bean golden mosaic virus (BGMV), 68
Bioethanol, 87, 99
Biotic stress, 161, 169, 187, 219
 miRNAs, 175
 siRNAs, 177
Biparental populations, 57, 60, 66, 70, 119
Bisulfite sequencing, 245
Brachypodium distachyon, 248
Brassica napus, 18, 40, 42, 147, 175, 247
Brassica rapa, 19, 146, 209, 218, 248
Breeding, 37
Bulk segregant analysis (BSA), 120

C

Cadmium (Cd), 171, 176, 218, 247
Capillary electrophoresis–mass spectrometry (CE-MS), 191, 193
Capsicum, 43
Cesium (Cs), 218
Chromatin, immunoprecipitation (ChIP), 252
 modification, 237
Chromium (Cr), 218
Chromomethylase3 (CMT3), 240
Chromosome segment substitution lines (CSSLs), 122
Clipper, 214
Clustered regulatory interspaced short palindromic repeats (CRISPR)/Cas systems (CRISPR-Cas9), 112, 161

- Cold acclimatization, 212
 Common bacterial blight (CBB), 68
 Composite interval mapping (CIM), 59, 282
 Compound identification, 195
 Copper (Cu), 176, 218, 269
 Copy number variations (CNVs), 20
Coriandrum sativum, 269
 Cotton, 19, 42, 65
 CRISPR/Cas, 140, 161
 activator (CRISPRa), 154
 interference (CRISPRi), 154
 Crops, improvement, 4, 20, 28, 37, 56, 129, 237, 255, 277, 283, 288
 sequencing/resequencing, 11
 Cucumber mosaic virus, 175
 Cyclic reversible termination (CRT), 15
- D**
- Data management, 20, 30, 277
 analysis & decision support tools (DMASTs), 8, 277
 Data mining, 187, 194
 Data processing, 194
 De-Bruijn-graph (DBG), 15
 Decision support tools, 1, 8, 286
 Differentially methylated regions (DMRs), 244
 Diseases, 43, 161, 243, 263, 269
 resistance, 43, 64, 86, 111, 121, 124, 170
 DNA, methylation, 237, 239, 255
 sequencing technologies, 11–19
 DNA methyl transferases (DNMTs), 244
 Doubled haploids (DH), 41, 63, 113, 133
 Double-digest RAD protocol (ddRAD), 40
Drosophila melanogaster, 117, 142–146, 239
 Drought, stress, 200
- E**
- Enzymatic mismatch cleavage (EMC), 146
 Enzymes, 263, 265, 268, 270
 Epigenetics, 165, 237, 239, 283
 Epigenome-wide association studies (EWAS), 248
 Epigenomics, 1, 7, 237, 243, 255, 290
 epiGWAS, 237
 Estimating genomics estimated breeding values (GEBVs), 279
 Ethylene-bis-dithiocarbamate (EBDC), 269
 Ethyl methanesulfonate (EMS) 140, 144
 Exon splicing enhancers (ESE), 166
- Exonic splicing silencers (ESSs), 166
 Expressed sequence tags (EST), 189
 Extremophiles, 213
- F**
- Fine-mapping, 110, 112, 122
 Flooding, 200, 208, 216
 Flowering time, regulation/control, 65, 101, 117–122, 126, 170, 241, 255
 Fourier–transform infrared (FT-IR), 191
 Fourier–transform mass spectrometry (FT-MS), 192
 Freezing tolerance, 212
 Functional genomics, 187, 189
 Functional mapping, 63
 Fungicides, 269
Fusarium graminearum, 219
Fusarium head blight (FHB), 68
- G**
- γ -amino butyrate (GABA), 202–221
 Gas chromatography–mass spectrometry (GC-MS), 187–211, 216, 219
 Gene discovery, 5–8, 133
 General linear model (GLM), 64
 Genetic gains, 277
 Genome-wide association studies (GWAS), 5, 65, 110, 116, 133
 Genomic breeding values (GEBVs), 277
 Genomics, 1, 156, 165, 278, 283, 288
 functional, 187, 189
 resources, 111
 Genomics-assisted breeding (GAB), 5
 Genomic selection (GS), 39, 62, 133, 279
 Genotyping by sequencing (GBS), 37, 39, 119
 Germplasm, 2, 5, 43, 68, 83, 91, 111, 115, 120, 132, 248, 283
 Glucose, 193, 197, 200–214, 218
 Glucosinolates, 218
 Glycolysis, 208, 209, 213, 215, 217
 Glycophytes, 214
 Glyoxylate cycle, 193
 Goat grass (*Aegilops tauschii*), 124
 Gold (Au), NPs, 268
- H**
- Haberla rhodopensis*, 201, 210
 Haplotypes, 28, 30, 112, 125

Heading date 3 (Hd3), 112
 Heavy metals, 176, 200, 209, 218, 251
 Heritability, 63, 69, 113–125, 148, 248
 Heterogeneous nuclear ribonucleoproteins (hnRNPs), 166
 Heterogeneous stock, 125
 High-density genotyping assays, 41, 129
 High resolution melt (HRM) analysis, 149
 High-throughput genotyping, 43, 71, 277
 platforms, 4, 283, 289
 High-throughput phenotyping, 68, 280, 288, 289
 High-throughput sequencing, 18, 20, 37, 165, 237, 252
 Histone deacetylases (HDACs), 245
 Homology-directed repair (HDR), 153
 HPLC, 145, 149, 190, 191
 with ultraviolet and photodiode array detection (LC/UV/PDA), 191
 Hybrid Sanger-NGS assemblies, 18
 Hydrogen peroxide, 215
 Hydroxycinnamic acids, 218
 Hypersensitive response (HR), 177

I

Identity by descent (IBD), 130, 131
 Immunoprecipitation, 252
 Insertion-deletions (InDels), 20
 Interval mapping (IM), 53, 56–61, 67, 131, 282
 Intronic splicing enhancers (ISEs), 166
 Intronic splicing silencers (ISSs), 166
 Iron (Fe), 218
 Isoenzymes, 87

J

Joint linkage-association mapping (JLAM), 5, 65

L

LC-NMR, direct infusion mass spectrometry (DIMS), 191
 Lead (Pb), 218
 Linkage disequilibrium (LD), 5, 30, 83, 86, 90, 110, 115, 129, 133, 283
 Liquid chromatography–electrochemistry–mass spectrometry (LC-EC-MS), 191

Liquid chromatography–mass spectrometry (LC-MS), 191

Lotus corniculatus, 214

Lotus crticus

Lotus japonicus, 146

Lupinus albus, 200, 201

M

Magnaporthe oryzae, 219

Magnolia kobus, 269

Maize (*Zea mays*), 18, 20, 28–31, 42, 63, 65, 101, 121, 124, 146, 176, 211, 241, 246, 248, 252, 286

Manganese (Mn), 218

Marker-assisted backcrossing (MABC), 284

Marker-assisted selection (MAS), 39, 56, 83, 86
 recurrent selection (MARS), 279

Marker-trait associations (MTAs), 53, 56

Mass spectrometry (MS), 187, 192

Meganucleases, 153

Meloidogyne incognita, 175

Memecylon edule, 269

Menta piperita, 269

Messenger ribonucleoprotein complexes (mRNPs), 172

Metabolomics, 1, 187

 limitations, 196
 plants, 197

Metal-graphene oxide (GO) NPs, 269

Methylation, 31, 170, 173, 196, 213, 237–255

Methylation-sensitive amplified polymorphism (MSAP), 249

Methyltransferases, 210, 239–242

Microorganisms, 153, 263, 269

Microwave-assisted extraction (MAE), 266

Molecular weight, 65

Mosses, 211

Multifounder populations, 123

Multiline cross inbred lines (MCILs), 65

Multiparent advanced generation intercross (MAGIC) population, 20, 65, 110, 117, 126, 133

Multiple interval mapping (MIM), 59, 282

Multiple-trait multiple-interval mapping (MTMIM), 62

Multivariate data analysis (MVDA), 194

Mutagenesis, 6, 153–155
 chemical, 139, 145

N

Nanomaterials (NMs), 8, 263, 264
 Nanoparticles (NP), 8, 263, 266
 metallic, 263, 270
 Natural antisense transcripts (NATs), 177
 Near isogenic lines (NILs), 122
 Nested-association mapping (NAM)
 population, 6, 20, 65, 109, 124, 133
 Next-generation sequencing (NGS), 6, 11, 14,
 140, 254
 NGS-only assemblies, 19
 Nickel (Ni), 176, 218
 Nitrogen, 216, 221
 deficiency/starvation, 176, 216
 NMR, 187, 190, 191, 203, 205, 209
 Nuclear factor Y (NF-Y), 176
 Nutrient deficiency, 216

O

Oryza sativa, 40, 87, 128, 147, 178, 202, 210,
 247, 248
 OSDREB2B, 169
 Overlap-layout-consensus (OLC), 15
 Oxford Nanopore (ONT) sequencing, 28
 Oxidative pentose phosphate pathways
 (OPPPs), 215
 Oxidative stress, 200, 215, 222

P

Pan-genomes, 30
 Parental selection, 97
 PCR, 15, 18, 39, 43, 146, 254
 Phenotyping, high-throughput, 68, 280, 288,
 289
 precision, 289
 Phenylpropanoids, 219
 Phosphorus, 216–218
 Phosphorylation, 196, 218, 241, 255, 267
Physcomitrella patens, 201, 211, 246, 253
 Phytohormones, 193
 Plant breeding, 56, 59, 61, 83, 87, 112, 161, 277
 Plant extracts, 263
 Plant splicingosomal proteins, 169
 Pleiotropy, 62
 Polyadenylation, alternative, 161, 166, 170
 Polyamines, 192, 200, 210–212, 219–221
 Populations, 2, 20, 40, 109, 111, 237, 245, 255
 biparental, 57, 60, 66, 70, 119

 crops, 37
 genetics, 85, 281–284
 mapping, 41, 53, 56, 83, 113, 124, 132
 multifounder, 123
 size, 6, 71, 114, 129, 132, 145
 structure, 31, 279
 sub-populations, 90
 Potassium, 216, 218
 Potato (*Solanum tuberosum*), 42, 65, 66, 87,
 192, 197
 Precision phenotyping, 289
 Presence-absence variants (PAVs), 20
 Proteomics, 1, 283
Pseudomonas syringae, 172, 175, 177
 PstavrRpt2 effector, 177
 Putrescine, 200–212, 218–221
 Pyrroline-5-carboxylate dehydrogenase
 (P5CDH), 178

Q

Quantitative disease resistance (QDR), 68
 Quantitative resistance loci (QRL), 68
 Quantitative trait loci (QTLs), 53, 56, 83, 86,
 133, 237
 linkage-based, 56

R

RADseq, 40
 Raffinose family oligosaccharides (RFOs), 222
 Rapid bulk inbreeding (RABID), 120
 Reactive oxygen species (ROS), 200
 Recombinant inbred lines (RIL), 41, 133
 advanced intercross lines (RIAILs), 65
 Recombination, 114
 Reduced representation sequencing (RRS), 37,
 40
 Replication, 28, 63, 66, 115, 148, 242
 Resequencing, 20, 30, 37, 41, 120, 255
 Restriction site-associated DNA (RAD), 39
 sequencing (RADseq), 37, 40
 Rice, 40, 87, 98, 112, 128, 147, 149, 168–171,
 178, 202, 210, 247, 248
 RNA, microRNAs (miRNAs), 161, 172
 single-guide RNA (sgRNA), 153
 single-stranded RNA (ssRNA), 242
 small interfering RNAs (siRNAs) 161, 172,
 242, 254
 RNA-directed DNA methylation (RdDM), 254

Root-knot nematode (RKN), 175
 Rubisco, 169

S

Salinity, 69, 176, 213, 220, 241, 251
 Salt stress, 213
 Sanger-NGS assemblies, 18
 Sanger-only assemblies, 17
 Seed dormancy, 170
 Sequencing by ligation (SBL), 14
 Sequencing by synthesis (SBS), 14
 Sequencing technologies, 11, 14, 27
 Ser/Arg-rich proteins (SRs), 166
 Serial analysis of gene expression (SAGE), 189
 Silver (Ag), NPs, 268
 Simple interval mapping (SIM), 58
 Single-locus analysis, 59
 Single-marker analysis (SMA), 57
 Single-molecule real-time (SMRT), 28
 Single nucleotide addition (SNA), 15
 Single nucleotide polymorphisms (SNPs), 20,
 37–41, 43, 112, 125, 143
 chips, 41
 Singlet oxygen, 215
 Skim sequencing, 40
 Small nuclear ribonucleoproteins (snRNPs),
 166
 Solid-phase micro-extraction (SPME), 265
 Soybean, 18, 20, 30, 65, 122, 146, 150, 168,
 176, 211, 216, 241, 248, 255
 Spliceosomal proteins, 169
 Splicing, alternative (AS), 161, 166, 168, 178
 RNA, 7, 161
 Spot blotch resistance, 87
 Starch, 87, 153, 193, 209–212, 217
 Streptomycin, 269
 Stress, abiotic, 7, 65, 118, 161, 169–178, 187,
 213, 219, 241, 243
 biotic, 161, 169, 172, 175, 187, 219
 combinations, 219
 drought, 200
 flooding, 200, 208, 216
 granules (SG), 172
 metabolomics, 191, 200
 oxidative, 200, 215, 222
 salt, 213
 temperature, 211
 STRUCTURE, 88–93
 Sub-populations, 90
 Sucrose, 193, 200–222
 Sugar alcohols, 192
 Sugarcane, 87, 104

Sugars, 192, 200, 268
 Sulfur, 176, 215–217, 219
 Sumolaytion, 255
 Supercritical fluid extraction (SFE), 265
 Superoxide, 215
 Systemic-acquired resistance (SAR), 269
 Systems biology, 187–190

T

Targeting induced local lesions in genomes
 (TILLING), 120
 Temperature stress, 203, 211
Thellungiella halophila, 213
 Thin layer chromatography (TLC), 191
 TILLING, in silico, 140, 151
 next-generation, 148
 Time-fixed mapping (TFM), 63
 Time-related mapping (TRM), 63
 TIR-NBS-LRR, 169
 Titanium dioxide, 267
 α -Tocopherol, 218
 Tomato, (*Lycopersicon esculentum*) 19, 43,
 66–69, 126, 127, 146, 150, 171, 175,
 189, 192, 197, 211, 213, 217, 246, 248,
 269
 Trait Analysis by aSSociation, Evolution, and
 Linkage (TASSEL), 283
 Trait dissection, 5
 Trait mapping, 5, 53, 86, 111, 113, 118, 125
 Transcription activator-like effector nucleases
 (TALENs), 153, 161, 164
 Transcriptomics, 1, 3, 7, 66, 161, 189, 210, 212,
 290

U

Ubiquitination, 239, 255
 Ultrasonication-assisted extraction (UAE), 265

V

Variant Call Format (VCF), 21
Verticillium longisporum, 172, 175
Vigna angularis, 28
 Visualization, 20, 24, 30, 170, 249, 277, 287

W

Water use efficiency (WUE), 200
 Wheat, 19, 41–43, 61, 66, 87, 99, 115, 124,
 139, 171, 178, 197, 211, 219, 286

Whole genome sequencing, 20, 37, 39, 60, 68,
142, 148, 151
 bisulfite (WGBS), 245
Whole genome shotgun strategy, 18

X

Xanthomonas oryzae, 219

Xanthomonas perforans, 270
Xanthomonas vesicatoria 269

Z

Zero-mode waveguide (ZMV), 28
Zinc (Zn), 216, 218
Zinc finger nucleases (ZFNs), 153